Master Thesis
Industrial Engineering and Management

Optimising internal benchmarking of mail delivery at PostNL

Svenja Gerdes

June 2017

UNIVERSITY OF TWENTE.

Author

Svenja Gerdes
University of Twente
Master program: Industrial Engineering & Management
Specialization: Production and Logistics Management
Graduation date: 16.06.2017
Student number: s1185667
E-Mail: s.gerdes@student.utwente.nl

SUPERVISORY COMMITTEE

Internal supervisors:
Dr. Ir. A. Al Hanbali
School of Management and Governance, University of Twente
Department Industrial Engineering and Business Information Systems

Dr. Ir. L.L.M. van der Wegen
School of Management and Governance, University of Twente
Department Industrial Engineering and Business Information Systems

External supervisor:
Ir. R. Veldman
Department Operational Strategy and Development, PostNL
The Hague, Netherlands

UNIVERSITY OF TWENTE.

Summary

Motivation and Research Goal

The recent decrease in mail volume resulted in a decline in income of PostNL Mail. Therefore, PostNL has to become more cost efficient in order to stay profitable within the mail sector. The delivery of mail is the highest element of expenditure within the mail sector of PostNL. Hence, there is an urgent need to increase the efficiency of delivery. PostNL introduced benchmarking between the delivery areas of the Netherlands to learn from each other and to exchange best practices between the delivery areas. However, the set-up of the benchmarking is lacking, in particular, the benchmarking model incorporates unsuitable performance measures and unrepresentative clusters, and the benchmarking sessions are improvable.
The goal of this research is to

develop an internal benchmarking model with adequate performance measures and clusters as a tool for process managers to determine best practices and performance gaps regarding mail delivery

New Benchmarking Model and Performance Measures

Benchmarking is a continuous process of analysing, comparing and self-improvement. For a successful benchmarking different steps have to be covered which we summarise in Figure 1. In this thesis we cover Step 1 to 8 with the main focus on Step 7, the clustering. The aim of clustering is to create high similarity between the areas within a cluster so that benchmarking between those areas will be fair. In particular we want to gain high similarity on factors which cannot be influenced by the management. In order to develop a cluster, we first have to define what we want to compare and the relevant factors, which is done in Step 6 by determining the most important performance measures and their influencing factors.

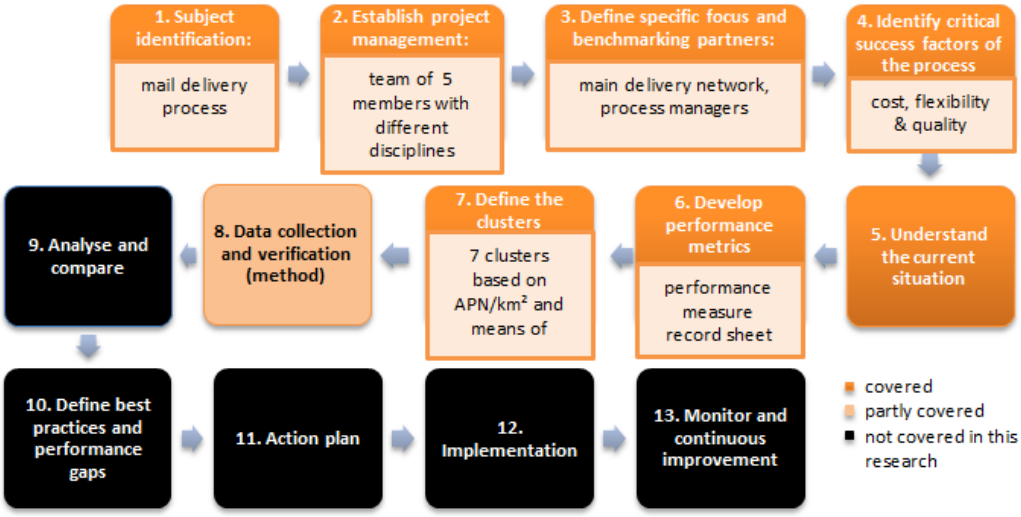


FIGURE 1: BENCHMARKING PROCESS AT POSTNL

We conclude that measuring performance productivity is most suitable for benchmarking. For a service like mail delivery, productivity should be measured in three aspects: the perceived service quality, the match between demand and supply and the output in relation to the input. In Figure 2 we suggest how PostNL can measure those aspects exactly.

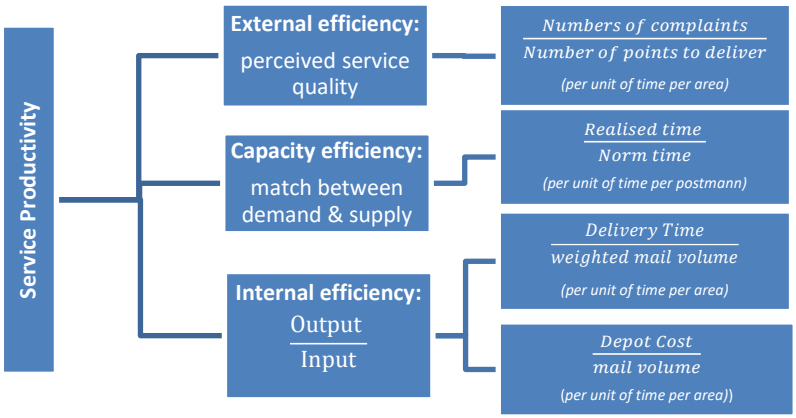


FIGURE 2: SERVICE PRODUCTIVITY AT POSTNL

As each performance measure is influenced by different factors, we suggest to develop a clustering per performance measure so that we can minimise the factors that have to be similar. Our research focuses on the performance measure delivery time per mail volume which relates input (time) towards output (delivered mail). Currently, the norm of delivery time is sorely based on historical data and the volume decrease and thus offers high potential room for improvement. Therefore, benchmarking the delivery time could identify this room for improvement and even provide strategies to do so.
The most critical factors influencing delivery time are interdrop, which determines the means of transportation, number of delivery points (APN), length of the minor-route per house and main-route. For the clustering we measure those factors per postcode 5 (PC5) areas. To enable a better comparability we measure the APN per km² and set the length of the main-route in relation to the APN (thus APN/m).

For calculating the mail volume within our performance measure, we advise to differentiate between the different kinds of mail and to give each kind a different weighting factor. The weighting factor should be based on the average time each kind of mail requires because this highly differs. For instance, ring-packages are much more work-intensive than letters.

Cluster Analysis

The testing framework is derived based on the methods suggested in literature and applied on all PC5 areas within the delivery area Utrecht as a representative sample for all delivery areas. For the clustering we tested different cases; first, we distinguished peak and off-peak days. Second, we have applied different combinations of factors for cluster analysis. For each case, we have applied different clustering techniques: automatic clustering (x-means), clustering techniques which require the number of clusters as input (k-means,) as well as a practical technique, which creates clusters without an optimisation criterion, but sorely based on the relationship between the factors.
The cluster outcome is evaluated regarding the compactness of a cluster and the separation between clusters. Therefore, we apply the Silhouette Coefficient, which has a range between 0 and 1, where one indicates highly compact and separated clusters. Furthermore, we apply the sum of squared error (SSE), which measures only the compactness and thus indicates the similarity between clusters. The closer the SSE is towards 0, the more compact the clusters are. However, no upper limit exists. We compare the outcome of the SSE and SC with the original clustering of the benchmarking model to evaluate the possible improvement (see Table 1Table 1). Finally, we have consulted an expert team to assess the reasonability of the cluster outcome.

TABLE 1: VALIDATION OF THE CLUSTER OUTCOME

Cases	Description	Validation criterion	x-means	k-means	practical	original
1p	Peak-day, given standardized (= equal weighted) attributes Interdrop, APN/m, minor-route	SC	0.52	0.42	-	-
		SSE	160	99	-	-
2p	Peak-day, given standardized attributes Interdrop, APN/m	SC	0.54	0.49	0.46	0.14
		SSE	57	41	52	88

The cluster outcome shows that PC5 areas that are grouped together on peak days are in the same group during off-peak days independent from the clustering technique and attribute set. Thus, we do not have to differentiate between peak and off-peak for benchmarking.

Based on the results of the validation criteria and the validation of the expert team we advise PostNL to apply the practical clustering for the benchmarking model, which defines clusters based on means of transportation and APN/km². The practical approach indicates a reasonable compactness and separation of the clusters with a SC of 0.46. Furthermore, it performs second best with its SSE of 52. Finally, the expert team selects the practical approach as the most reasonable and realistic one off all approaches.
The practical approach improves the original clustering by around 40% in its cluster compactness. The original clustering performs with a SC of 0.14 poorly on the overall compactness and separation and results given the SSE of 88 in the lowest similarity between objects within one cluster.

Recommendations for PostNL

This research shows that there are no highly distinctive clusters for delivery time and makes clear that only some factors (interdrop and APN/km²) incorporate cluster tendencies, which are considered in the final clustering, although there are many others factors influencing the delivery time. Therefore, benchmarking delivery time can help to find reasons for difference in the time needed, however will not be highly precise. Based on this study we can conclude that the current information infrastructure is quite elaborated, yet. PostNL should make use of it and develop and apply a norm model that allows a more precise estimation for the required mail delivery time.
Anyway, to ensure that benchmarking delivers value independent of the performance measure that is benchmarked, it is essential that PostNL covers all steps of the benchmarking process: Besides comparing performances and discussing best practices, process managers should also define an action plan clarifying the best practices, their implementation, evaluation and monitoring.

After successfully implementing the internal benchmarking model, we also advise to develop an external benchmarking with other postal companies in order to determine global best practices. While the main national competitor Sandd might not be willing to share information, companies across the border like Denmark or Belgium might be willing to cooperate as they face the same problem, but in a different market.

Furthermore, the mail process at PostNL contains different sub-processes including collection, sorting & preparation and mail delivery. Current bottlenecks of the mail process are the links between the sub-processes. For instance, a delay in the delivery of mail to depots often hampers a smooth mail process. The current management and control system focuses on each sub-process separately, however to determine specific problem areas between sub-processes we advise to implement a monitoring and control system for those links as well.
Finally, to incorporate the complexity of the mail process within a measurement system and still provide a clear view on the performance levels with their individual performance factors, we advise to use an Analytic Hierarchy Process (AHP) based methodology, independent whether it is for a control model or a benchmarking model. This method places the performance in a hierarchical order, directly showing how the performance measures are interlinked.

Preface

This paper is not only the result of my graduate internship at PostNL, but also symbolises the end of my studies at University of Twente. I had wonderful, interesting and challenging five years in the Netherlands, and the last half year in The Hague at PostNL formed a successful conclusion of it. During my seven months of graduate internship at the department Operational Strategy and Development of PostNL I had the possibility to observe and to take part in the process of mail delivery. They gave me room for self-initiative and the support to conduct my research for improving the benchmarking model. By applying not only the learnt theories of my study Industrial Engineering and Management but also new acquired once from the academic literature to practice, I gained a lot of new experiences and knowhow.

I want to thank all those who helped me carrying out this research. I am grateful for my colleges at PostNL who made me feel being part of the team from day one, and I am thankful for their openness for collaboration, discussions and exchange of (working) experience. Furthermore, I want to thank Helen Verschoor who gave the initial assignment and enabled a good start at PostNL. Together with her, Laurie ter Maart, Jan Weijers, Ronald Veldman and Rogier van den Brink I had a great project team to whom I could present my interim results and who gave me feedback from multiple perspectives. My special thanks go to Ronald Veldman, my supervisor at PostNL, despite of his busy schedule, he was always available to give me helpful guidelines. His valuable suggestions and critical questions helped me not only to reflect and improve the choices I made for the thesis but also taught me a lot for my personal development.
Next to the people from PostNL, I am thankful for the assistance of my supervisor Ahmad al Hanbali from the University of Twente. He was always open to my ideas, but also helped me to stay on track by providing me constructive feedback. I would also like to thank Leo van der Wegen, the second member of my graduation committee, who helped me to improve my thesis by evaluating and commenting it.

My grateful thanks also extended to my family for their support and motivation. Finally, I would like to acknowledge the encouragement provided by my friends throughout my studies.

Enschede, June 2017

Svenja Gerdes

Contents

List of Abbreviations and Definitions

Abbreviation	Full word (Dutch term)	Definition/ Description
	case	Cases are derived based on different combinations of subsets and scenarios. For instance Case 1p considers the attributes of Subset 1 given peak days (Scenario SP)
	mailbox packages	packages that do not exceed 380x265x32mm and a weight of 2kg
	parameter	factors that define a system and determine its behaviour. Those factors set conditions of its operation, but cannot be directly influenced by the user.
	ring package (bel pakje)	package that requires a signature or does not fit through the mailbox with a maximum dimension of 380x265x125mm
	cluster attribute	an attribute (also referred to as feature, variable, dimension, component, factor within the academic literature) based on which objects are assigned to clusters
	interdrop	distance on the main-route between two succeeding delivery points
(K)PI	(Key) Performance Indicator	Performance targets which "focus on the outputs of an organisation"(Johnson, Whittington, & Scholes, 2011, p. 446)
aHC	agglomerative hierarchical Clustering	one kind of hierarchical clustering
APN	delivery point (afgiftepunt)	a physical address for mail delivery, registered at the Base Register of PostNL (BRPP)
BAG	register of addresses and buildings	this registration is administrated by the Dutch Government
BG	delivery area (bezorg gebied)	area which is managed by one process manager
BRPP	base register of PostNL	system in which PostNL stores all delivery points (including information such as their addresses, minor-route)
CSF	Critical Success Factor	"areas in which results, if they are satusfactory, will ensure successful competitive performance for the organisation" (Rockert, 1979, p. 85)
dima	distance of main-route	distance that a postman has to walk on the public street to cover all delivery points given a tour or an area
dimi	distance of minor-route	distance from the main-route to the mail box of a delivery point
HC	hit-chance	estimated percentage of delivery points of a tour that actually receives mail
K		the number of clusters within a clustering

MJ dashboard	manage and justify dashboard	the dashboard contains different key performance indicator and is used by the control department and managers to control and evaluate the performance
NVR	Network Volume Registration	system in which PostNL stores the amount of Mail per delivery point per day
pbz	postman (postbezorger)	Employees of a postal company who deliver mail to given addresses
PM	performance measure	“as a metric used to quantify the efficiency and/or effectiveness of an action” (Neely, Gregory, & Platts, 1995, p. 1229).
PC5 area	areas given postcode 5	areas given the first five digits of the postcode
RBV	resource-based view	According to this view a firm can create a sustainable competitive advantage based on the core competences of its resources
RU, RO	run-up, run-off	distance between the start/end point of the main-route and the depot
S&P	sorting and preparation	One part of the overall mail process. The first part of the mail process is collection, subsequently sorting and preparation and finally mail delivery.
S1	Subset 1	subset of attributes incorporating interdrop, minor-route and APN/m
S2	Subset 2	subset of attributes incorporating interdrop and APN/m
SI	Scenario Infrastructure	attributes that are evaluated independent from the mail volume
SO	Scenario off-peak	attributes that are evaluated given the off-peak mail volume
SP	Scenario peak	attributes that are evaluated given the peak mail volume
SSE	Sum of the Squared Error	a validation criterion for measuring the compactness of a cluster
USO	Universal Service Obligations (Universele Postdienst voorwaarden)	obligations for the Mail department of PostNL given by the Dutch government
WTR	base working time (Werktijd regeling)	estimated time by the process optimisation department for a given delivery tour

1. Introduction to Research

In our society there is a strong growth of digital alternatives substituting physical mail leading to a sharp decrease of transactions of physical mail. Within five years, the physical mail volume decreased by one third in the Netherlands (ACM, 2016). Still, our expectations of the physical mail service are constantly increasing: national mail sent today should be delivered tomorrow, important mail and packages should be traceable, and if a package cannot be delivered, we want to pick it up nearby.

However, the sharp decrease of physical mail transactions leads to less income for postal operators. Therefore, the main challenge is to stay profitable within this mail sector. This research supports PostNL, the Dutch mail, parcel and e-commerce corporation to improve the efficiency of the mail delivery. We propose a benchmarking model to identify performance gaps as well as best practices for the delivery of national mail.

In this chapter, we introduce PostNL and the problem background of the national mail delivery. Subsequently, in Section 1.3, we discuss the possible causes and consequences in order to identify the core problem. Afterwards, in Section 1.4, we define our research goal and set the scope of this project in Section 1.5. Furthermore, we outline the plan of approach on how we can reach our goal and based on that define the structure of this thesis, which we present in Section 1.6. Finally, in Section 1.7, we determine the research question and sub-questions which need to be answered in order to solve the research problem.

1.1. Introduction to PostNL

For over 200 years PostNL has been responsible for the delivery of mail in the Netherlands. It started in 1799, when the Dutch government introduced the first national mail company. The postal law, established in 1807, ensured that they were the only company allowed to collect, to transport and to deliver mail.

In 1989, the national mail company changed to a private one called PTT Post (*Staatsbedrijf der Posterijen, Telegrafie en Telefonie)*. To prevent a decrease in quality of the mail service, the government defined Universal Service Obligations (USO, in Dutch: Universele Postdienst voorwaarden) for PTT Post, which are still valid for PostNL nowadays. Those obligations include delivering 95% of the mail the next day – five days a week – and provide sufficient letterboxes and post offices (PostNL, 2016d).
During the 21st century the Dutch post market was liberalized and competitors like Sandd B.V. entered the market. In order to stay competitive, the company had to redesign itself and changed within 20 years three times the corporate identity as well as its brand name (1. PTT Post, 2. TPG Post, 3. TNT Post).
In 2011, TNT Post decided to demerge into two independent companies: TNT Express, which focuses on the international courier delivery service, and PostNL, which focuses on the mail and parcel service (see Figure 1.1).

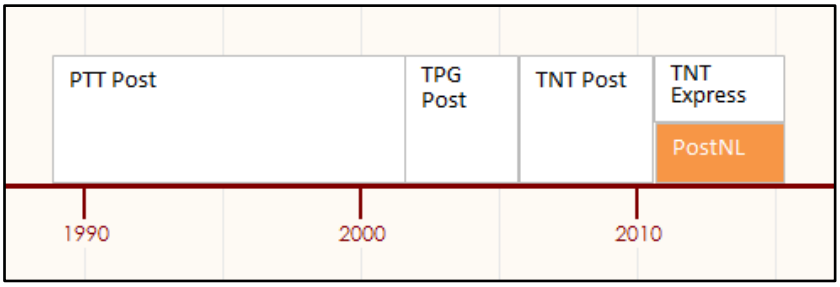


FIGURE 1.1: DEVELOPMENT SINCE THE PRIVATISATION

Currently PostNL consists of three main business segments; Mail in the Netherlands, Parcels in the Benelux and International. Those three segments are managed separately; the international segment deals with their subsidiaries Spring Global Delivery Solutions, Nexive in Italy and Postcon in Germany. The mail and parcel segment have their own collection and sorting process as well as their own delivery-network (PostNL, 2016a).

This research focuses on the business segment Mail in the Netherlands. Within this segment we can differentiate between business and consumer mail. Business mail concerns business to business (B2B) and business to consumer (B2C) mail delivery, while consumer mail implies consumer to consumer mail delivery. 96% of the mail volume is business mail; the delivery terms (e.g., costs, frequency of collection and delivery) depend on the individual contract set between the business and PostNL. Consumer mail, which is only 4%, requires a complex network as PostNL has to fulfil the USO terms which obligates them to deliver to every physical address under the same conditions (e.g., 5 days a week, delivery by the house entrance). In order to remain profitable with such a low mail volume, it is essential to have an efficient delivery network.

1.2. Problem Background

Observing PostNL over the last five years, we can see a constant increase in the international and in the parcel business segment. However, the volume of mail delivery in the Netherlands is constantly decreasing (see Figure 1.2). While PostNL had 3,777 Million items of mail delivery in 2011, it declined to 2,401 Million in 2015, which implies a 36% decrease. This has an effect on the revenue stream which was declining from 2,439€ Million in 2011 to 1,961€ Million in 2015, thus showing a 24% decrease (PostNL, 2016b).

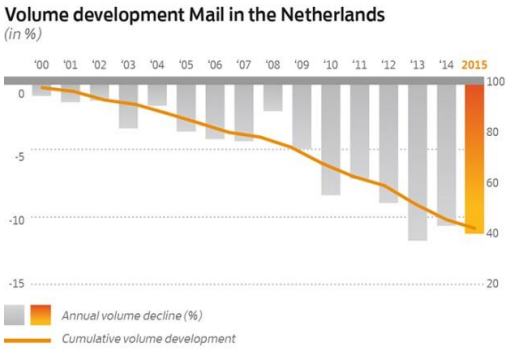


FIGURE 1.2: VOLUME DEVELOPMENT MAIL IN THE NETHERLANDS FROM POSTNL (2016A)

This development is mainly due to the growth of digital alternatives substituting physical mail, resulting in decreasing transactions of physical mail. This leads to two major challenges for PostNL. Firstly, PostNL wants to keep a high level of service quality and customer satisfaction in this segment as those two factors are their key differentiators and PostNL has to fulfil the obligations of the USO. However, it is hard to keep a high service level and to stay profitable, while the income is decreasing. Secondly, the shrinking market of the mail delivery results in an intensified competition between postal operators. In order to cope with those challenges PostNL has to create a more efficient mail process to face the pressure on pricing by competitors (PostNL, 2016a).

The mail process in the Netherlands can be separated into 3 parts: collection, sorting & preparation and delivery. The main expenses of the mail process lie in the mail delivery, which also contains the most room for improvement. While the process of sorting and preparation is fully standardized, it is difficult to define accurate norms and measures to control the efficiency of the mail delivery process. In order to find best practices and to determine performance gaps for the mail delivery, the control department of PostNL designed and introduced a benchmarking model in 2014. Before that each management level has had and still has dashboards with (key-) performance indicators to control the performance. However, those dashboards primarily concentrate on financial reporting. The main

goal of the benchmarking model is to develop an alternative analysis and evaluation tool for the mail delivery which is rather focused on the operational performance outcome. The idea is that non-financial performance measures enable a deeper perspective by shifting the focus towards the operational drivers that enable cost reductions, giving the managers better indications on what to improve to realize a more efficient mail delivery process.

The performance measures of the benchmarking model are determined for each delivery area within the Netherlands. Every three months the control department organizes a benchmarking session to compare and to discuss the performance outcomes with the delivery process managers of each delivery region. However, the benchmarking sessions do not work as expected; the performance of the different delivery areas is hard to compare and to evaluate as a delivery area is not homogeneous, but contains different geographical areas. For instance, there is one overall performance measure for the delivery area Groningen although it contains cities as well as rural areas. Hence, the measurements per delivery area can contain high variation. The process managers keep struggling to interpret and evaluate the given benchmarking data instead of finding the best practices. Furthermore, the control department noticed declining motivation of the process managers to join the benchmarking sessions. In the following section, we identify the problem of the current benchmarking model by determining the roots of the problem and the consequences it has.

1.3. Problem Identification

The benchmarking model is designed by the control department. It determined performance measures and the form of presentation after a small consultation with process managers of different delivery areas. The clustering for the benchmarking model is based on the already existing area division defined by management; they divided the Netherlands into 28 delivery areas each led by one process manager. The process manager is responsible for around 10 team leaders within that area, each managing around 110 postmen (pbz) (see Figure 1.3).

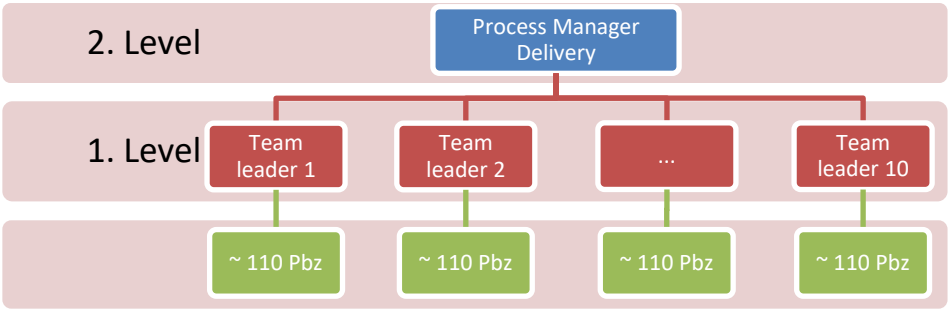


FIGURE 1.3: MANAGEMENT OF ONE DELIVERY AREA

The control department divided the delivery areas into five clusters with the aim to minimize the differences between them; big cities, highly urban, medium urban, lower urban and rural. The clustering is based on delivery points (APN) per km^2 of each delivery area (see Table 1.1), which we refer to as cluster attribute; a delivery point can be defined as a physical address for mail delivery, registered at the Base Register of PostNL (BRPP). Furthermore they differentiate within a delivery area between car, scooter or remaining deliveries (incl. bike, e-bike, foot).

TABLE 1.1: CURRENT CLUSTERING OF THE DELIVERY AREAS

Cluster	APN/km2	Name
A	> 1000	Big cities
B	500 - 1000	Highly urban
C	300 - 500	Medium urban
D	175 - 300	Lower urban
E	< 175	Rural

Solely clustering based on APN/km^2 per delivery area results in two main problems:

1. **High variation within the delivery area (cluster objects)**
   The current objects that are clustered form the delivery areas. This clustering allows no differentiation within a delivery area although it contains different geographical areas, which vary highly on the number of delivery points per km² (APN/km$^{2)}$ and require different delivery strategies. Given the example of delivery area Groningen (see Figure 1.4), we can see that it contains rural areas like Pieterburen with a low APN/km^2 where a scooter or car delivery would be the most efficient. However, it also contains cities like Groningen with a high APN/km^2, which can be best delivered by bike or foot. Consequently, there is a high variation within the delivery area which is not considered within the performance measures.

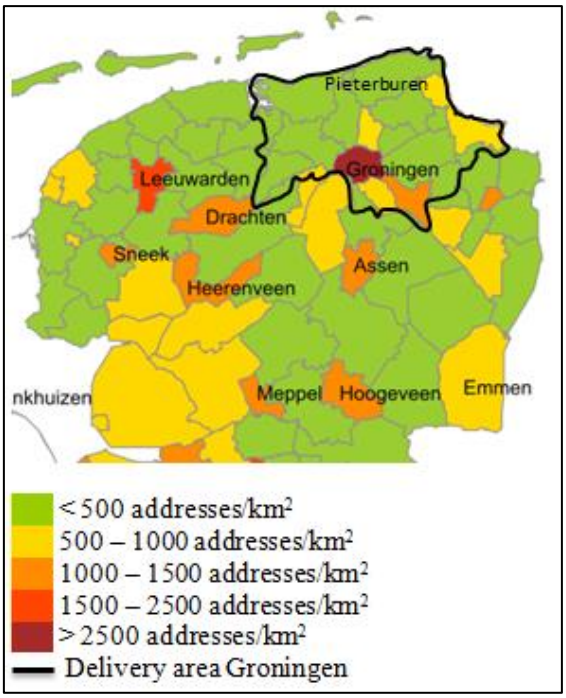


FIGURE 1.4: DELIVERY AREAS OF NORTH-EAST NETHERLANDS

2. **Low similarity within one cluster**
   The benchmarking model compares different performance measures. Each measure depends on different factors, both influenceable (e.g. means of transportation) as well as factors, that characterize the area, but cannot be influenced by PostNL (e.g. APN/km^2, mail volume). From now on we refer to those non-influenceable factors as parameters. To ensure a fair comparison, we should compare only those areas with similar parameters being relevant for the performance measure. However, until now there is only one cluster attribute for all the different performance measures. Consequently, depending on the performance measure, the degree of similarity within one cluster can vary.

During the benchmarking sessions we noticed the resulting problem: Due to the high variation within a delivery area, a process manager cannot identify root causes of his high performance scores and consequently cannot give any advice to the other managers. Furthermore, instead of trying to find best practices by comparing their performance with the other delivery areas, they spend most of the time in arguing why a benchmarking was not possible due to the differences between the delivery areas although they are in the same cluster.

Additionally, process managers often do not see the value of evaluating certain performance measures and rather spend their time in discussing other topics during the benchmarking session.

Finally, they struggle to interpret the graphics as the measurements are too complex or not well explained. Hence, performance measures as well as their form of presentation are not suitable for motivating the process managers to compare and discuss their performance.

Overall, the process managers are losing their interest in the benchmarking sessions, mainly because an efficient comparison is not possible due to the heterogeneous clustering, but also because the discussed performance measures are too complex. Consequently joining the benchmarking session does not add much value for them.

All those problems can be summarized by the overall problem statement:

The current benchmarking of the national mail delivery is inadequate.

The core problem of an inadequate benchmarking can be divided into three sub-problems; The first sub-problem is the composition of the benchmarking, which does not meet the needs of the stakeholder. The incorporated performance measures are too many and on a too low processing level (rather data than information), which makes it harder to interpret the model. The second sub-problem is the technique used for clustering, because the current clustering does not meet the aim of creating high similarity within as well as a high difference between clusters (Tan, Steinbach, & Kumar, 2005b). The final sub-problem is the execution of the benchmarking during the benchmarking sessions, because those sessions do not reach the goal of triggering discussions on performance improvements. (see Figure 1.5)

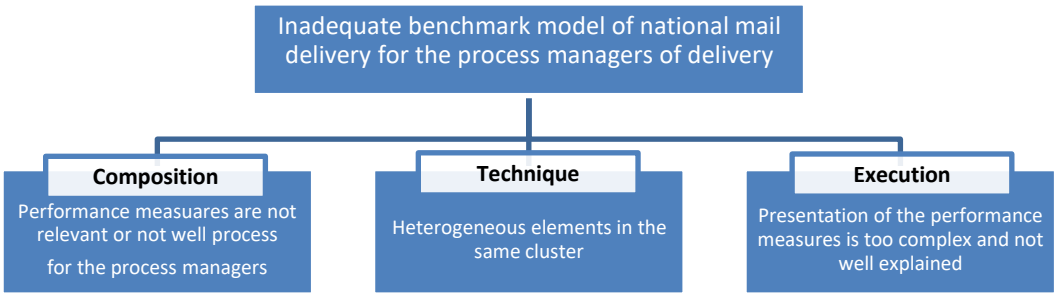


FIGURE 1.5: AN OVERVIEW OF THE PROBLEMS

1.4. Research Goal

The mail business segment of PostNL is under high pressure to increase the efficiency of mail delivery. Benchmarking is a useful tool to identify best practices and to improve the process. It can be conducted internally as well as externally. Internal benchmarking compares the performance within an organisation, while external benchmarking compares the organisation with other organisations, for instance direct competitors (Anand & Kodali, 2008).

Internal benchmarking can provide significant benefits, but only, if the organisation meets the following criteria (Southard & Parente, 2007):

- *similar processes*: The mail delivery process at PostNL is similar throughout the Netherlands
- *adaptable techniques*: The techniques used for the mail delivery process of PostNL can easily be adapted because firstly the techniques are not too complex and secondly change and adaption are well established within the company's culture.
- *superior processes*: The performance of delivery areas differs highly. In the current benchmarking there is a big gap between top and bottom score of the performance measures which gives much room for improvement.
- *available performance metrics*: PostNL has an overload of data of each delivery area. However, it is not clear yet how it can be used to operate the national mail delivery more efficiently.
- *transferable practices*: Due to the similar processes within the Dutch mail network many practices could be transferred easily, but it is hard to determine the best practice for the national mail delivery.

Overall, the mail business segment satisfies the criteria for an internal benchmarking model. Nevertheless, the current benchmarking model does not reach the goal of improving the mail delivery process due to the inadequate set-up as seen in the problem identification (1.3.). We want to solve this problem by determining useful performance measures, homogeneous clusters and an appropriate way of presenting the benchmarking data with the overall aim to improve the benchmarking. Therefore, the research goal is as follows:

Develop an internal benchmarking model with adequate performance measures and clusters as a tool for process managers to determine best practices and performance gaps in mail delivery.

1.5. Research Scope and Limitations

As stated above, sorting and preparation are fully standardised which allows sufficient control for and measurement of efficiency. However, it is difficult to define accurate norms for the mail delivery process. Even though the mail delivery process has the highest expenses of the whole mail-process, it still lacks adequate tools for controlling and improving the efficiency. Therefore, this research aims to develop a benchmarking model for comparing and improving the mail delivery performance of PostNL.

Mail delivery consists of two different networks which are managed separately. The main mail delivery network contains all the physical addresses of the Netherlands, around 8,000,000 delivery points, and follows the USO terms. This network is highly complex and difficult to control due to the high quantity and distribution of delivery points. The lack of control makes it more difficult to identify problems in the mail delivery process, resulting in lower efficiency.
The other network concerns mostly mail delivery to parties which set a special delivery contract with PostNL. Currently, with around 20,000 parties its network size is only 0.25% of the size of the main network. This small size makes it quite comprehensive and easier to control than the main network. Therefore, the benchmarking model of this research focuses on the main mail delivery network to enable better control and performance improvement. Starting point of the network is when the postman leaves the depot for mail-deliveries by foot, bike, e-bike and scooter or the HUB for mail-deliveries by car (see Figure 1.6).

FIGURE 1.6: THE MAIL-PROCESS

PostNL delivers mostly addressed mail via the main mail delivery network, except for Saturday, when unaddressed mail such as flyers for advertisement is distributed as well. For our research we exclude unaddressed mail to ensure comparability between all delivery days.

Target group of this project are process managers of the delivery areas, hence we limit it to performance measures which are within their management scope, meaning everything downwards and including the second level (see Figure 1.3).

Moreover, we will determine clusters solely based on performance measures defined for the benchmarking model. Hence, clustering is only based on parameters that influence those performance measures.

Due to time constraint, we set different priorities for solving the sub-problems (see Figure 1.5). Biggest challenge of the core problem is to gain knowledge on clustering techniques and to apply them to the benchmarking model as currently PostNL lacks competences in this area. Besides, the construction and execution sub-problem is a less challenging as it is more an issue of critically revising current performance measures and form of presentation. Thus, highest value can be added by solving the technical sub-problem, and this will be the main focus of this research. However, clustering depends on performance measures, therefore we revise them in this paper, but keep the research as limited as possible. Considering the execution sub-problem, we provide tips that should be considered, but do not conduct a detailed research on it.

Given the time frame, we are not able to build and to implement the benchmarking model in the company framework. However, we give suggestions for the design and setup of the model. Furthermore, we aim to determine a clustering technique that can be applied uniformly independent from the performance measure. As mentioned, we cluster on parameters influencing a performance measure. Therefore, if we can show that this technique is applicable to a performance measure we can expect the same result for the remaining ones. Hence, taking into account the time constraint, validation and testing will be limited to a clustering for one performance measure.

Finally, we apply and test cluster analysis only on a representative sample instead of the whole Netherlands, because data collection and computation time of conducting a cluster analysis for the whole Netherlands would exceed our time frame. The exact sample will be defined later on in our test framework (see Section 6.1).

1.6. Plan of Approach

The problem we are going to solve is an action problem. An action problem can be defined as “a perceived (by the problem-owner) discrepancy between norm and reality” (Heerkens, 2004, p. 2). The norm set by the control department is that by using the benchmarking model process managers could determine best practices to improve the mail delivery process. However, in reality process managers are not able to identify best practices as the benchmarking model is not adequately designed. One well-known method to solve an action problem is the Managerial Problem Solving Method (MPSM). Using this method we will be able to eliminate the perceived discrepancy between norm and reality. To validate the findings of each step, a project group of managers from different departments relevant for the benchmarking model is formed, to which the findings are presented. MPSM includes the following steps:

1: identifying the problem
2: planning the problem-solving process
3: analysing the problem
4: generating alternative solutions
5: choosing a solution
6: implementing the solution
7: evaluating the solution

In Chapter 1, we have already covered the first two steps: We have clarified the problem background and context to identify the core problem. In particular, we have analysed the development of mail delivery within PostNL and briefly the benchmarking model and sessions. For the problem-solving process, we have defined the project goal and scope (see Section 1.4 and 1.5).

In order to perform the remaining steps, different information and knowledge are required. In the following section we derive the steps and information that is needed to successfully execute the

MPSM and to reach our research goal. We do this by defining research questions and sequence on solving them which determines the structure of our research project.

1.7. Required Information and Research Questions

In the following section we define research questions and sub questions. By answering those questions, we gain all required information and knowledge which enable us to solve the action problem.

Context Analysis

We first need to understand the organisational and the operational structure of the mail delivery process to determine requirements, limitations and constraints of the benchmarking model. Therefore, we answer the following questions in Chapter 2:

1. How is the national mail delivery process of PostNL organised?
    a. What are the steps of the mail delivery process at PostNL?
    b. Who is responsible for which part of the process? (organisational chart)
    c. How is the information and control structure within the mail delivery process? (information flow chart)

We will conduct a secondary source data collection in form of a content analysis of the information within the system of PostNL (quantitative as well as qualitative). Furthermore, we will collect primary data by observing the mail delivery on different days with different means of transportation in order to examine differences in the areas and to develop the mail delivery process scheme. Moreover, we will observe team leaders, process managers and their team meetings to design the organisational chart and the information flow chart.

Evaluation of Current Model

In Chapter 3, we analyse and evaluate the current benchmarking to identify its strengths and weaknesses. Therefore, we answer the following research questions:

2. How is the current benchmarking organised?
    a. What is the goal of the current benchmarking model and which performance measures are defined?
    b. Which parties are involved in the benchmarking and how?
    c. What are interests and needs of the stakeholders?
    d. To what extend are those needs satisfied?
    e. What are bottlenecks and problems of the current benchmarking model?

Those questions can be answered by secondary source data collection in form of a content analysis of the old benchmarking session presentations and the current benchmarking model. Furthermore, we obtain information by observing the benchmarking sessions. Finally, we conduct a stakeholder analysis to determine stakeholders' interests and needs. To do so, we conduct a qualitative research as it is more sensitive and provides more freedom for exploration than a quantitative research. One of the most used qualitative methods is conducting interviews (Babbie, 2009). We use a semi-structured interview for our stakeholder analysis, because we do not exactly know their interests and needs, but still want to enable a comparison between the interviews. As mentioned in the research scope and limitations (see Section 1.5), we limit this research as much as possible. Therefore, we conduct semi-structured interviews with a representative sample of stakeholders of the benchmarking.

Literature Review
To gain knowledge on developing an adequate benchmarking model, we answer the following questions by conducting an academic literature review. We mainly use Web of Science and Scopus as those two search engines have the broadest and largest interdisciplinary databases for Science, Technology, Engineering and Medicine. By using two engines there will be a broader range of possible good and suitable articles. Furthermore, to ensure high quality of the articles, we select based on the numbers of citations, the year of publishing and the journal using the Scimago Journal & Country Rank. The findings will be presented in Chapter 4.

3. How can an adequate benchmarking model for the mail delivery be designed according to academic literature?
    a. How does academic literature define an adequate benchmarking model?
    b. What method can be used to design an efficient benchmarking model?
    c. What are criteria for good performance measures?
    d. Which performance measures are suitable for a delivery process?
    e. What solution approaches for clustering exist in the literature?

Develop performance measures and derive cluster attributes for the benchmarking model of the mail delivery process
In Chapter 5, we will combine findings of the previous questions to develop performance measures and clusters for the new benchmarking model.

4. Which suitable performance measures can be defined for the national mail delivery at PostNL?
    a. How can the criteria based on the literature review as well as the interest and needs of the stakeholders (Question 2.c. & 3.c) be applied to the mail delivery process of PostNL?
    b. How can those performance measures be defined and measured?
    c. Which parameters influence the performance measures and can be used as cluster attributes?
    d. Is the current information structure sufficient for calculating the performance measures?

In this step, we use a qualitative data collection by semi-structured interviews over data availability within the control and IT-department. To identify which parameters influence the performance measures, we analyse the environment of mail delivery by observation and interviewing (semi-structured) experts within PostNL on mail delivery and making use of already existing models within PostNL.

Perform and evaluate the cluster analysis for the benchmarking model
In Chapter 6 we will define and apply the test framework for conducting a cluster analysis given the performance measure derived and selected in Chapter 5. For this framework we apply the methods identified during the literature review (Chapter 4) by adapting them to the specific problem context. Therefore, we answer following questions:

5. What should be the clustering for the benchmarking model?
    a. Which elements should be incorporated in the test framework to ensure a high quality clustering?
    b. Which clustering approach performs best given the performance measure for the benchmarking model?

Practical implication and suggestions for the implementation
By defining performance measures in Chapter 5 and clusters in Chapter 6, we have the main input for the benchmarking model. In Chapter 7 we assess the fit of the clustering with the managerial

structure and clarify possible implications for the quality of the benchmarking. Finally, we will present a proto-type benchmarking model, including the criteria derived during the literature review and the interests and needs of the stakeholder defined in Chapter 3.

6. How should the new benchmarking model for PostNL be designed?
    a. What are possible implications when implementing the performance measure and clustering in the benchmarking model?
    b. How should the new benchmarking model for PostNL be designed?

The thesis will end with an overall conclusion and recommendation in Chapter 8.

2. Mail Delivery Process at PostNL

In this chapter we will give an overview of the mail delivery process. We will describe each step and outline main factors of the process that might vary (Section 2.1). Furthermore, in Section 2.2., we will present the organisational structure of the delivery process and in Section 2.3 the inherent information flow that is used to control and to evaluate the mail delivery process. Finally, in 2.4, we will summarize their effect on our three categories of technique, composition and execution.

2.1. Mail Delivery Process

PostNL delivers to more than 8 million addresses in the whole Netherlands 5 days a week, with 26,500 postmen (PostNL, 2016d). The mail volume is higher on Tuesday, Thursday and Saturday, the so called peak days, and lower on Wednesday and Friday, the off-peak days. There are different types of mail, which have to be delivered by the postman. PostNL differentiates between addressed mail, which include letters, transactions, direct mail, mailbox packages as well as ring-packages, and unaddressed mail such as flyers for advertisement, which are only delivered on Saturday.
In the following we briefly describe the steps of the mail delivery process considering only addressed mail as defined in our scope. For a detailed description we refer to Appendix I.

Each delivery tour is assigned to a certain depot, except of delivery tours by cars, those pick up the mail at HUBs. The mail to the HUBs is delivered before 9 a.m., because postmen that deliver by car, have to start their tour at 9.30 a.m. Mail to depots is delivered at different time slots, one at 11 a.m. and the other one at 1 p.m. Deliveries from there have no mandatory starting time, but have to be finished before 6 p.m.

The postman loads the bags of mail at the depot or HUB on his means of transportation, which can be the post boy for foot deliveries, bike, e-bike, scooter or car (see Appendix II). In case that not all bags fit on the means of transportation, the postman has to reload during his tour. From the depot or HUB, the postman goes or drives to his first delivery point, which is the starting point of the delivery tour (see Figure 2.1, no.7). The distance between depot and starting point, so called run-up (see Figure 2.1, no. 1), can vary per tour.

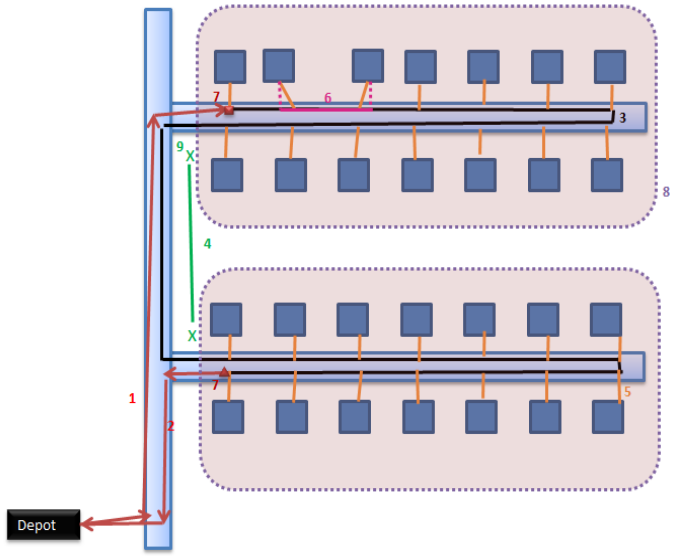
Every postman has a certain route for his delivery tour (see Figure 2.1, no.3), that he has to follow, called main-route (see Figure 2.1, no. 3). The length of the main-route varies per delivery tour, but has a limitation per means of transportation.

Some delivery tours, mostly those ones by bike, contain sub-tours (see Figure 2.1, no. 8), implying that the postman has to park (see Figure 2.1, no. 9) and step off his/her means of transportation, take the bundle of mail for that sub-tour and walk one round for the mail delivery. If the delivery tour has no sub-tours, the postman can stay with his means of transportation. The sequence on delivering the addresses and sub-tours are specified for all tours.

The mailboxes are not always reachable from the street. Often, for instance, if houses have front yards, the postman has to walk a minor-route from the street to the mailbox (see Figure 2.1, no. 5). During the walking the postman grabs addressed mail out of the bundle to directly place it in the mailbox. If it is not possible to put all the mail in the mailbox or if an item requires a signature, the postman rings the bell of that address and tries to hand it out to the resident. If no one opens, the postman tries to contact neighbours so that they can forward the mail later on. If three neighbours do not open, the postman will bring the mail to a specified retailer at the end of his/ her delivery-tour. To inform the resident, the postman puts a standardised form informing about the location of the post (neighbours or retailer) into his mailbox.

If the sub-tour is finished, the postman goes back to his/her means of transportation and rides/drives to the next delivery point or to the parking spot of the next sub-tour until he/she reaches the end of the main-route.

During the tour it might not be possible to deliver all mail. In addition to the reason already stated above, there are two more. Firstly, the mail is sorted in the wrong tour, meaning that the address is not within that certain delivery tour, but in another. Secondly, the address does no longer exist, or mail is not accepted by the resident. All this mail has to be equipped with a sticker informing about the cause. If possible, the mail has to be put into the public mailbox of PostNL by the postman. Otherwise the postman has to bring it to a specified retailer. Car deliveries form an exception; they neither have to seek for a public mail box nor a retailer, because they have to return undelivered mail to the HUB.
After taking care of the undeliverable mail, the postman can go home directly and return the bags in their next shift, except if they have a post boy, e-bike or car as those have to be returned to the depot or HUB before going home.



No.	Term
1	run-up
2	run-off
3	main-route
4	connection route
5	minor-route
6	interdrop
7	start- and endpoint
8	sub-tour
9	parking spot

FIGURE 2.1: MAIL DELIVERY TOUR

Differences between Delivery Tours
Although the process of mail delivery at PostNL is the same through the Netherlands, qualifying a good internal benchmarking (Southard & Parente, 2007), there are factors influencing the execution. By joining and observing different delivery tours on different days, we recognised varying factors, which should be considered when developing homogeneous clusters:

1. **mail volume**: As stated at the beginning, PostNL has peak and off-peak days. Consequently, the number of mail items is varying. If the volume is increasing, the average number of items per delivery point is increasing as well as the number of delivery points that have to be delivered.
2. **numbers of actual addresses to deliver**: A delivery tour always contains the same delivery addresses, but not all addresses receive mail. The number of houses that actually receive mail is varying per day and per delivery tour.
3. **distance between delivery points**: The distance between delivery points varies. In rural areas we noticed a higher distance than in urban areas, which is mainly due to the density of households or the type of building (e.g. row houses, detached houses or blocks of flats). PostNL differentiates between the distance from the main-route to the delivery point, the so called minor-route (see Figure 2.1, no. 5) and the distance between delivery points at the main-route, the so called interdrop (see Figure 2.1, no. 6).
4. **total travel distance**: While the main-route of the tour is fixed, in case of undeliverable mail the postmen might have to cover additional distances to the retailer and/ or to the public mailbox of PostNL as well.

5. **means of transportation**: There are different means of transportation (foot, bike, e-bike, scooter or car). They are set by the process optimisation department based on the average interdrop, speed and hourly charge. Hence, we exactly know which means of transportation is used per tour.

2.2. Management and Control System

In this section we give an overview on the management and controlling of the mail delivery process. We have a look at the information exchange between the different management levels with a focus on downwards and including the second management level (see Figure 1.3).

PostNL differentiates between four management levels within the operations of mail delivery. The 4th and highest one is the national level managed by the director of preparation & delivery. The 3rd one is the regional level. PostNL divided the Netherlands into six regions - northeast, central, northwest, west, southwest and southeast - and allocated them over three managers, the so called region managers of delivery. Each region consists of different delivery areas, which represents the 2nd level. Each delivery area is managed by a process manager of delivery. The 1st and lowest managerial level is the team leader. The delivery tours of one delivery area are divided on around eight to ten team leaders, who manage the postmen that walk those tours.

Information and Control 2nd – 3rd Level

The control department is responsible for monitoring level two to four. To do so it determines Key Performance Indicators (KPIs), which are performance targets focusing on the output of an organisation (Johnson et al., 2011), each year and creates different performance dashboards based on them. Those are used by process managers to manage their area, but also to control the performance as the KPI realisation is always compared to the predefined budget. There are two dashboards, one that is updated weekly, called MJ week dashboard, and another one updated monthly, MJ month dashboard (see Figure 2.2). Both have similar KPIs (see Table 2.1). However, whereas MJ week dashboards are independent from each other and aim to give a screen shot of the current performance, the month dashboard incorporates all information until the current month. By that the month dashboard shows trends and can be compared to targets of the year. The results of the dashboards are discussed within managing & justifying (MJ) meetings. During those meetings process managers have to justify their performance results to their manager and a controller.

In Table 2.1 we summarize the KPIs which show the frequency of measurement as well as their weighting factor. Whereas three KPIs, engagement of region, engagement of own delivery area and culture of own delivery area, are measured via a questionnaire that every employee has to fill out once a year, all the remaining PIs are included into the MJ month dashboard and partly into the MJ week dashboard. Both dashboards contain KPIs for costs, quality & control and employees. The employee performance does not fluctuate highly during a week, and therefore the control department decided to include only one KPI (absentee) of the category employees in the MJ week dashboard. Furthermore, as mentioned above, unaddressed mail is only delivered once per week, as one measurement is not sufficient to evaluate the performance, it is not discussed in the MJ week dashboard, but only in the MJ month dashboard.
Finally, every month process managers gain an overall performance score on cost, quality & control and employees based on their realisation and the weighting factor given for each KPI.

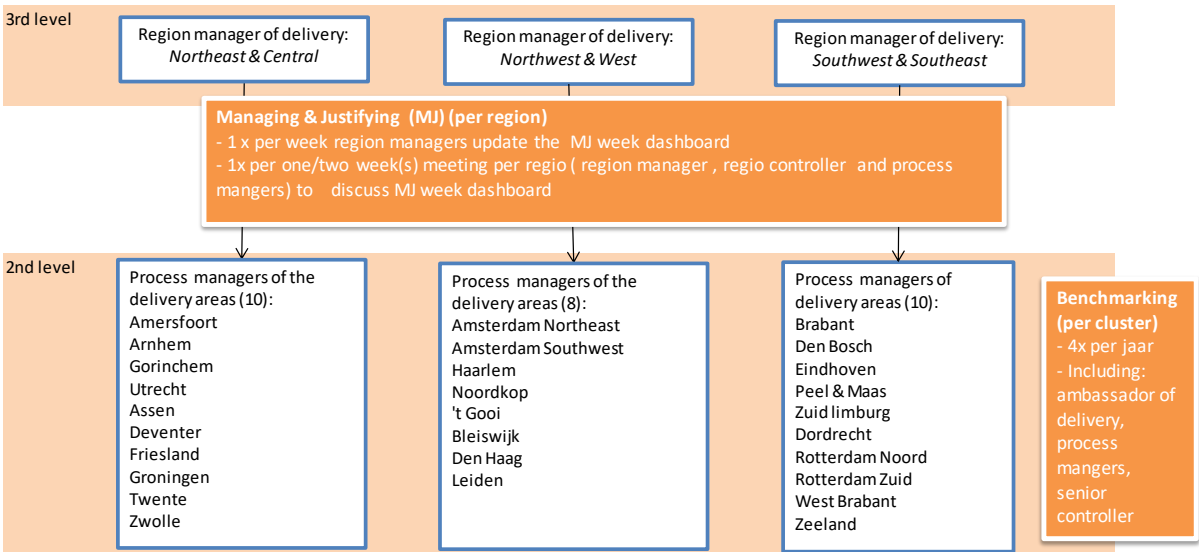


FIGURE 2.2: ORGANISATION AND CONTROL OF THE DELIVERY MANAGEMENT (2ND-3RD LEVEL)

TABLE 2.1: PERFORMANCE INDICATORS OF THE MJ DASHBOARD (#=MEASURED IN NUMBERS)

KPI	Unit of Measurement	Weight-ing	Measurement Frequency
Costs			
Total cost of the delivery area incl. sorting and preparation (S&P)	€	1	monthly
Total cost of the own area	€	1	weekly, monthly
Quality & Complaints			
Delivery time mailbox packages of region	% on time delivered	1	monthly
Delivery time mailbox of own delivery area (incl. S&P)	% on time delivered	1	weekly, monthly
Delivery time to big customers of own delivery area (incl. S&P)	% on time delivered	1	monthly
Number of business complains own area	#	1	weekly, monthly
Number of private complains own area	#	1	weekly, monthly
Quality unaddressed mail own area	% of mail with sufficient quality	1	monthly
Quality delivery own area	% of mail with sufficient quality	1	weekly, monthly
PDCA (plan–do–check–act) measurement own area	% reached of an overall score	1	weekly, monthly
Number of bags not delivered	#	1	weekly, monthly
Employees			
Total absenteeism % of own area	%	1	weekly, monthly
Engagement of region	Score 1-100	0.5	yearly
Engagement of own area	Score 1-100	0.5	yearly
Culture of own area	Score 1-100	1	yearly
Contract-size PBZ of own area	#	1	monthly
Volunteer mobility of own area	#	0.5	monthly
Pension outflow in own area	#	0.5	monthly
Outflow of region	Score 1-100	1	yearly
Work accidents of own area	#	1	yearly

Used in the MJ dashboard of the 1st and 2nd management level

Benchmarking Model

Whereas MJ meetings are mandatory, benchmarking sessions are voluntarily. Data for the benchmarking model is collected and updated automatically. The region manager northwest & west took the responsibility as ambassador for delivery to manage the benchmarking together with the senior controller. Every three months they organise a benchmarking session for each cluster, visualising parts of the benchmarking in different graphs and discussing the performance measures of the different delivery areas with the process managers. Contrary to MJ meetings, it is not a control tool and does not contain targets that they have to fulfil.

The benchmarking model contains four different categories for measuring the performance:

1. **logistic**: total number of delivery points, number of depots and total as well as average rent of the depots
2. **employment**: number of employees per task (e.g. administration, pbz's on peak and off-peak day, pbz's for sorting the inner bags into the shelves at the depots, pbz's at car and scooter delivery) and their attendance (e.g. % and costs of absenteeism, days of vacation, trainings)
3. **composition of employees**: age distribution of pbz's and groups of employees (e.g. % of short-term contract, car drivers, Saturday workers)
4. **labour agreement**: wages, hours and total costs per employee group as well as negative hours of the working time

The information is measured and presented per quarter. One can select quarter and region, delivery area or cluster to determine which results are shown in the model. If you select based on the delivery area, the model automatically displays all delivery areas of the related cluster.

Benchmarking Session

Every quarter there ideally is a benchmarking session per cluster organised by the senior controller and the ambassador for delivery. The control department summarises results and determines abnormalities of the quarter before. It decides whether the next session involves a more general benchmarking or zooms in on one specific performance measure. Results and abnormities are graphically presented with some underlying questions in a PowerPoint presentation, which the control department sent in advance to the process managers so that they can prepare themselves and think about possible explanations.

The benchmarking session starts with discussing the action points of the previous session and the topic of the current session. Subsequently, the group is divided into subgroups mostly randomly to discuss the presentation and the underlying questions. Afterwards their findings are discussed with the whole group, and action points, if required, are formulated.

In reality, one of the four sessions is skipped due to seasonal issues (e.g. high absenteeism during summer). Furthermore, the attendance is sometimes quite low. To ensure a reasonable group size the sessions of different clusters are merged into one. Finally, the control department tried to involve the process managers in determining topics of the next benchmarking session by giving them the possibility to send ideas regarding the topic via mail. However, the answer rate was extremely low.

A more detailed analysis and evaluation of the benchmarking is given in Chapter 3.

Information and Control 1st – 2nd Level

The KPIs used in the MJ dashboard at the 2nd line are incorporated in the MJ dashboard for the 1st line (see orange marked KPIs in Table 2.1). The process manager organises a meeting with the team leaders to discuss the KPIs once a month.

Besides the MJ dashboard, the process manager uses the integral dashboard and the customer compliance system to manage the team leaders on a daily and weekly basis (see Figure 2.3). The integral dashboard shows the performance indicator delivery-time of the pbz's per means of

transportation on a team leader level (scooter, car, bike & foot), the number of sick calls and the number of customer visits by the team leader are presented as well. The customer compliance system shows all complaints from customers that are received by the customer service. Information about the date, location and kind of complaint are given and directly linked to the responsible team leader. Similar to the MJ dashboard, KPIs not fulfilling the targets require a justification from the team leader.

2.3. Available Information on the Mail Delivery Performance

In the current system process managers gain information over the actual mail delivery from five sources (see Figure 2.3).

The first source is the process optimisation department, which provides an estimation of how long the delivery tour should take, the so called base working time (WTR time). As already mentioned, the accuracy is low as it relies on historical data and the development in mail volume.

The second source is the pbz, which reports the required time per deliver tour via the smart phone app 'PostNL my work' by indicating the short- or overtime.

The third source is the team leader, who visits around 3 pbz's to observe and to assess their performance. Next to that, they gain information on the delivery quality and customer satisfaction by visiting around 3 customers a day. However, considering that one team leader has around 110 pbz it is hard to collect enough information for each pbz in an acceptable timeframe.

The fourth source of information is the customer complaints received by the customer service of PostNL. Those complaints can be linked to the delivery-tour and pbz responsible. The cause of complaint is analysed by the team leader who has to contact the customer. However, this complaint system only gives an indication on the quality of delivery, because not all accrued complains are submitted to the customer service.

The fifth source is an external company, hired by PostNL, which measures the quality of delivery based on a survey on around 70000 customers. Furthermore, the company controls the total delivery time from the sender until the receiver. Those measures are conducted on a postcode 4 level, meaning that the findings can be seen per city district, but not on a delivery tour level. Hence, results cannot be linked to the performance of a pbz.
Overall, we can see that there exists no direct information on the performance of delivery. It either depends on the reliability of the pbz, the perception of the team leader or the received customer complaints, or the information is given on such a high level, that it cannot be directly linked to the performance within a delivery tour.

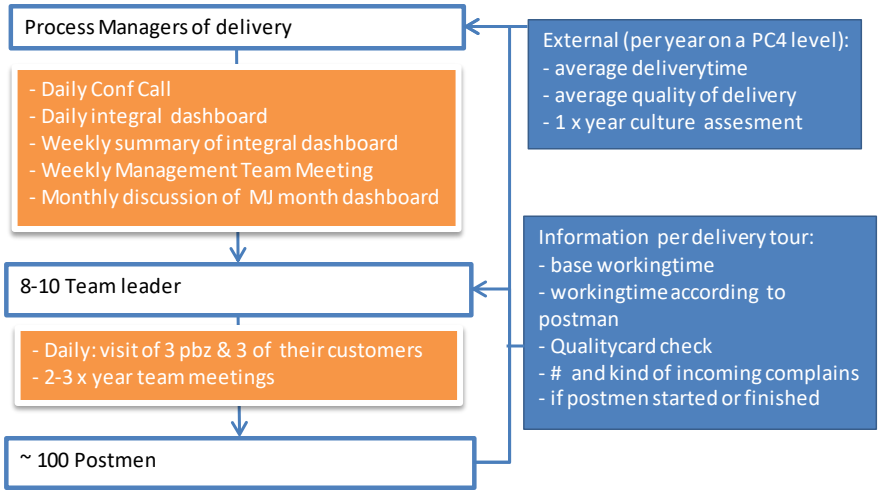


FIGURE 2.3: INFORMATION AND CONTROL SYSTEM DOWNWARDS THE 2ND LINE

2.4. Conclusion

Overall we can come to following conclusions:

Composition – Dashboard and Performance Measures

The MJ and integral dashboards function as a control tool focusing on performance output. However, in order to improve the process one should also relate the input towards the output, which current performance measures of the benchmarking model fail to do. It includes some input factors (e.g. number of pbz or depot), but none of them is related to the output.

Furthermore, we can see that the control and information system is standardised for the whole Netherlands, but MJ dashboards are discussed independently per region and per delivery area. Thus, a national comparison would be possible, but is not used except for the benchmarking model.

Technique - Clustering

Firstly, as seen in the MJ dashboards, KPIs do not have to be measured in the same frequency, but the frequency depends on the fluctuations of the (K) PI, which can be applied to the benchmarking model as well.

Secondly, we can conclude that PostNL has a clear division of responsibility within the production department. In order to ensure a good implementation we should consider the management structure for our clustering as well. However, sorely conducting clustering based on the management structure can lead to heterogeneous clusters as shown in the current benchmarking model. Therefore, for the final clustering, we have to determine an approach which results in homogeneous groups but also complies the managerial structure.

Execution – Organisation and Management of the Benchmarking Session

The control circuit of MJ meetings is clear and well structured; meetings have a fixed date, and the tasks are clear: process managers have to prepare a justification. In contrast, planning and organisation of the benchmarking are quite vague. Meetings are planned short dated, and, as it is voluntary, the attendance per meeting is quite varying. Finally, all MJ and integral dashboards are directly related to each other, and therefore each manager is motivated to score well in all of them. However, the benchmarking is not linked to the overall control system and thus less prioritised by process managers.

3. Analysis and Evaluation of the current Benchmarking

In Section 2.2 we have already described the current benchmarking model and the benchmarking session. In the next step we will analyse stakeholders of the benchmarking and their perception of the benchmarking model and sessions. Subsequently we will define the gap of the current benchmarking based on the findings of the stakeholder analysis and the current management and control system.

3.1. Stakeholders of the Benchmarking Model and their Interests

Only by determining who has a stake in the benchmarking and understanding their interests and needs we can design an adequate benchmarking model. Therefore, we will firstly conduct a stakeholder analysis by using the interest-power grid (Slack, Chambers, & Johnston, 2010), which helps us to position stakeholders in the benchmarking project. Secondly, we will conduct semi-structured interviews with the stakeholders to better define their needs and interest.

Positioning the Stakeholders

One of the most used methods within analytical categorisation of stakeholders is the power-interest matrix. Stakeholders are plotted into that matrix depending on their degree of interest and influence. This helps specifying to what extent a stakeholder wants and should be engaged in the regarding project. (Reed et al., 2009)

We follow the power-interest grid presented in Slack et al. (2010), which suggests for each kind of stakeholder a certain strategy: Stakeholder with low power and less interested groups should be only monitored without high engagement. Low powerful, but interested groups should be adequately informed to ensure that no major issues are arising. If the stakeholder has high power, it is important to keep them satisfied. While powerful stakeholders with low interest do not demand active involvement, highly interested once expect fully engagement (Slack et al., 2010).

For the benchmarking model we can identify five main stakeholders: the process manager who uses the benchmarking model and organises the mail delivery process within his delivery area, the people who organise the benchmarking (ambassador of the delivery, the senior controller responsible for delivery), the team leader who has to control postmen and to implement the wishes of the process manager, the postman who executes the delivery and finally the senior advisor process optimisation which joins the sessions to give consultation.

In Figure 3.1 we plot the stakeholders of the benchmarking in the power-interest grid. The detailed analysis of their power and interest can be found in Appendix III.

Given the plotting and strategies suggested by Slack et al. (2010) we can make following conclusions on the involvement of stakeholders. **Postmen** have low interest as well as low power, therefore an involvement of them in the benchmarking is unnecessary, but as they are the last chain of the mail delivery process, we will interview them briefly to realise a deeper understanding of the actual mail delivery. In contrast, **team leaders** are interested in the benchmarking, but do not have the power to change the delivery process without the agreement of the process manager. Therefore, we should keep them informed about the results of the benchmarking, but not actively involve them. Until now, this is working well as process managers pass important and relevant findings of the benchmarking sessions to their team leaders. **Process managers** have high power on the mail delivery process, however they have more interest in MJ dashboards than the benchmarking. Thus, they have medium interest. Therefore, it is important to keep them satisfied, but only involve them if they gain value out of it. To ensure that they are satisfied with the benchmarking model, we will interview a representative sample of process managers to determine their interests and needs. The **senior controller** has high interest as changes in the model or session lies within her responsibility. She conducts the data collection and analysis of the benchmarking model and has high power by doing so. Therefore, we should highly involve her and ensure her satisfaction as without her agreement we could not make any changes. The **ambassador of delivery** is responsible for the execution of the

benchmarking sessions and has a powerful position in the structure of the benchmarking model as well as in the mail delivery process. Thus, we should ensure that he is satisfied and engage him especially in areas related to the execution of the benchmarking. Finally, with his passive role as an advisor the **senior process manager of optimisation** has low power for the benchmarking. However, as he is responsible for improving the overall quality of the mail delivery process, his interest is high. Therefore, we will inform him about the decisions and ask him for his advice, but it is not critical to satisfy him.

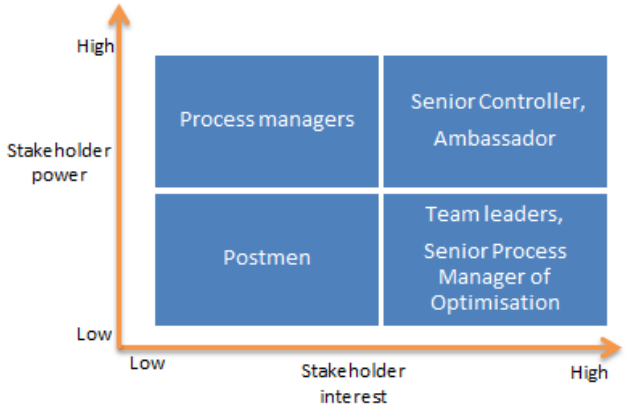


FIGURE 3.1: THE POWER-INTEREST GRID APPLIED TO THE STAKEHOLDERS OF THE BENCHMARKING MODEL

Semi-structured Interviews

As we have seen in the positioning of stakeholder, it is critical to satisfy the powerful stakeholders. To determine their needs and interests, concerning the delivery process, the benchmarking model including the clustering as well as the benchmarking sessions, we conduct semi-structured interviews. In the following we will present the approach and results. The overall conclusion is given in Section 3.2.

In the power-interest grip (see Figure 3.1) we have seen that even though the postmen, the Senior Process Manager of Optimisation and the team leaders have no power, they play an essential part in the delivery process. Thus, they can give us some insight into factors that might influence the delivery performance. Therefore, we will interview them as well, but will keep it as limited as possible.

Approach

We will interview the senior controller and ambassador, but due to the time constraint, we will not be able to interview all 28 process managers. We have selected a representative sample with the following requirements. From a managerial perspective, we chose one process manager from every region as every region manager has its own management style and view on the mail delivery process. Furthermore, to assess the impact of the homogeneity within a cluster, we select a delivery area which is mostly homogeneous given the current cluster attribute (APN/km²) and one with highly varying densities. In order to ensure sufficient experience the process manager has to have at least 3 years' experience, preferable in different delivery areas. Overall, we have a sample size of three process managers that fulfil our requirements.

Considering team leaders and postmen, we set one year of work experience as a requirement. To gain a high variety we select them from different delivery areas and postmen with different means of transportation (car, scooter, and foot/bike).

Overall, we conduct 12 semi-structured interviews, 3 with process managers (which are 11% of all process managers), 3 with team leaders, 3 with postmen as well as one with the senior controller, the senior process manager of optimisation and with the ambassador. We have designed two nearly identical interview templates, one for stakeholders who are joining the benchmarking and one for

the others. For detailed information about the template set-up and the execution of the interview see Appendix IV.

Results

The main purpose of the interviews is to explore three things: firstly, how learning and improving are incorporated in the current system, secondly, what the interests and needs of the stakeholders regarding the mail delivery process are, finally, how the stakeholders assess the current benchmarking (clustering, performance measures & session) and how they would improve it.

Learning and Improving within the Mail Delivery Process

In all cases, except one, the learning and improvement were triggered by a problem that occurred. This shows us that there is more a reactive approach instead of critically reflecting and actively searching for improvements. Most of the time, after proven to be successful, the learnings are shared with colleagues and their manager during a team meeting. However, how and if the colleagues apply it was not really emphasised by the stakeholders. This might indicate that it is rather an isolated "everyone on their own"-culture than a group responsibility for improving the mail delivery together.

The Mail Delivery Performance

There is a coherent perception over the mail delivery performance through all levels: All respondents say that the mail delivery was working well and that the *quality* was good: PostNL has built a fine-tuned network that reaches every place in the Netherlands and the mail is delivered to the right address by a representative postman on time. We can see that those criteria are mainly those ones that they have to fulfil according to the USO, and hence, it is no surprise that it works well.

There are three main areas that require according to the stakeholders improvement. One is the *link between sorting & preparation and delivery* as delay is not always communicated, or the incoming mail is not sorted well. Second is the lack of *transparency* within the process due to missing data which leads to an evaluation rather based on intuition than on facts. Finally, indicated by half of the interviews, is the poor *connection to the postmen* as the communication from both sides could be more. However, the last point highly depends on the management style of the process manager and team leader, which can explain that the other half of our sample indicated a good communication.

Overall, we can see that no one sees need for improving the mail delivery process. As we can see from their learning and improving, stakeholders mostly realise areas of improvement if problems occur. However, the transparency is missing to directly realise problems in the delivery.

The Performance Evaluation

We have summarised the proposed performance measures per management level in Table 3.1. There are four main categories which all stakeholders would evaluate: quality, costs, employees and customer satisfaction. However, not every category is evenly emphasised by each management level. We can see that costs are only pointed out by the higher management levels, whereas team leaders or postmen rather emphasise soft factors like behaviour towards customers or commitment of the postmen.

Furthermore, process managers only named factors that are already part of the MJ dashboards. Only stakeholders outside the delivery department named new factors like the numbers of mail delivered per hour or average costs per mail item. This indicates that either the process managers are satisfied with the current performance measures and do not see the need to measure something else or they do not think outside the MJ dashboards on their own.

TABLE 3.1: PERFORMANCE MEASURES ACCORDING TO THE STAKEHOLDERS

	Specialists (ambassador, control, optimisation)	Process Manager	Team Leader	Postmen
Quality	Correct delivery Complaints Same as in MJ dashboard	Complaints	Complaints On time delivery with no damage at the correct address	Complaints On time delivery at the correct address
Efficiency	mail delivered per hour			
Cost	Average cost per mail item	Cost per means of transportation		
Employees	Commitment Employee Turnover	Absenteeism	Commitment Overtime Flexibility	Commitment
Customer satisfaction/ perception		Customers' perception	Appearance of pbz Behaviour towards customers	Appearance of pbz Behaviour towards customers

Clustering

The most named parameters that can vary within areas are the distance between delivery points, the layout of the house (e.g. flats, industry building), the location of the mailbox and the arrangement of house numbers within a street. For the whole list of factors see Appendix V.
All the participants of the benchmarking, except one, are satisfied with the current cluster attribute (APN/km²), but it should be on a lower level as there is still a high difference within one delivery area. The majority named big cities, medium cities and villages as a sufficient clustering. The process manager disagreeing with the current cluster attribute emphasises that clustering is never possible in his opinion as there always are various influencing parameters that differ. This goes along with the team leaders and the postmen who indicate that each delivery area has its own characteristics. This shows the need for a separate clustering for each performance measure, because by this we can instead of including all parameters for the clustering limit ourselves to only those that influence the performance measure.

Benchmarking Model

There is a high consensus on the advantages and the disadvantages between the participants of the benchmarking.
Main disadvantage according to the stakeholders is the complexity of the model as it contains too many performance measures which hamper an easy exploration and analysis. Furthermore, the data is not always correct. For instance, control receives only information about the total depot costs per region and estimates the costs per delivery area, although the exact costs per depot are available. Finally, the benchmarking model is not often used but only opened if the control department sends a reminding mail mentioning the upcoming session.

Advantage of the model is the nationwide comparison which gives the process managers more insights into their colleagues. Moreover, the model does not set any budgets like the MJ dashboards which allows to sorely focus on the realisation. Finally, which was named by half of the participants, they like that it is partly linked to the performance indicators of the MJ dashboards.

Overall, we can conclude that the process managers are interested in seeing how they perform compared to others. Furthermore, the benchmarking model is not a convenient tool for the process managers as it takes too much time and effort to analyse the large amount of information and to draw conclusions. Hence, the new benchmarking model should have a limited number of performance measures. Finally, process managers are more interested in performance measures which are linked to the MJ dashboards. Other performance measures are less valued and

emphasised by them. Thus, if we want to gain commitment of process managers for the benchmarking model, it is advisable to select performance measures that are to some extent connected to the MJ dashboards.

Benchmarking Sessions

There is not only a high coherency in the assessment of the benchmarking model, but also in the benchmarking sessions.
All the process managers value to come together, face-to-face, and to have an open discussion with their colleagues. Furthermore, they highly appreciate to receive underlining questions from the control department as a starting point for discussions. Overall, stakeholders indicated that those discussions helped to think critically and deeper about their performance, which was considered less during MJ meetings.

However, benchmarking sessions are quite loose; every participant prefers to make it mandatory with clearly defining the input and the output of such a session. Furthermore, more responsibility should be given to the process managers by, for instance, letting each process manager by turns prepare the sessions. Moreover, even though everyone wants to learn, there is still a quite tense atmosphere, where people feel to justify themselves, rather than an open discussion. Finally, some are not satisfied with the criteria for forming subgroups. The idea is to form it based on the cluster. However, it is often just formed randomly.

All stakeholders indicate that 4 sessions per year would be sufficient to discuss all the developments, except for one, who would rather hold the benchmarking once a month connected to the monthly MJ meeting.

3.2. Conclusion: Gap between current Situation and the Goal

In Section 2.4 we have already presented conclusions based on the analysis of the current management and the control structure, which are defined in Gap 3 and 7 (see Figure 3.3). Together with the findings of the stakeholder analysis we have defined the overall gap between the current situation and the goal of benchmarking and summarised them in Figure 3.3. We can see that some of the findings of the stakeholder analysis support our problem definition of 1.3 and/or the conclusions in Section 2.4. This includes the high number of performance measures, the highly varying parameters per delivery tour and the missing guidelines for the sessions (see Figure 3.3, Gap 1, 5 & 7). In order to solve those gaps, we defined different requirements for the composition, technique and execution of the benchmarking.

First, based on the analysis of the mail delivery and the stakeholders, we can conclude that the input and the output are quite different in each area. Therefore, if you want to compare the performance fairly, it is advisable to compare the ratio between those areas. Moreover, given the ratio (input/output), it is much easier to evaluate the efficiency of the performance (Gap 3) and also to minimise the number of performance measures (Gap 1). Providing the possibility to split this higher processed performance measure up in sub measures, we still enable process managers to analyse the performance in more detail if required (Requirement 1).
Secondly, to ensure that performance measures are calculated correctly, we have to determine if the information is available or obtainable on the required level before setting the final performance measures (Requirement 2).
Thirdly, to gain homogeneity within a cluster, we do not only have to define smaller cluster objects than the delivery area, but also have to limit the high number of varying parameters affecting the performances (Gap 5) by designing a separate clustering for each performance measure (Requirement 3).
Finally, to improve the organisation of the benchmarking (Gap 7) we have to define clear guidelines on tasks for and organisation of the benchmarking sessions (Requirement 4). The organisation

should also be to some extant the responsibility of the process managers, leading to an overall higher participation.

Goal	Learning from and exchanging best practices with each other		
	Composition	**Technique**	**Execution**
Requirements	1. Present a higher level of analysis by combining performance measures. 2. Before setting the final performance measures define information requirements and check if it is realisable	3. To ensure low variance within a cluster define cluster objects not only on a lower level, but also for each performance measure separately.	4. Define clear guidelines for the benchmarking 5. Assign the organisation of the session to some extant to process managers
Problems ■ based on delivery process analysis ■ based on stakeholder analysis ■ based on both	1. Too many performance measures 2. Effortful exploration and analysis of performance measures 3. No relation/ratio between performance output and input that could indicate the efficiency of a practice 4. Wrong data: Information or information infrastructure not available	5. Many varying parameters at the lowest level of the mail delivery process	6. No mandatory delegation of responsibility to process managers 7. No clear guidelines
Current situation	Rather open discussions than concrete benchmarking		

FIGURE 3.2: GAP BETWEEN GOAL AND CURRENT SITUATION

4. Literature Review

To answer Research Question 3, *How can an adequate benchmarking model for the mail delivery be designed according to the academic literature?*, we will review academic literature on benchmarking, performance measures and clustering in this chapter. Based on this we will define criteria and methods which we can apply to develop the benchmarking model of mail delivery process at PostNL.

4.1. Benchmarking

Benchmarking is a well-known tool for organisational learning, widely practiced (Dattakumar & Jagadeesh, 2003; Voss, Åhlström, & Blackmon, 1997; Yasin, 2002) and successfully applied to various functional areas (Dattakumar & Jagadeesh, 2003). By measuring and comparing business practices within or between organisations, one can determine best practices, which, after having understood and adapted them, can lead to an overall organisational improvement (Anand & Kodali, 2008; Camp, 1989; Drew, 1997; Southard & Parente, 2007).

One of the pioneers in benchmarking is Camp (1989), who showed how Xerox Logistic and Distribution has significantly improved their productivity and efficiency by applying competitive benchmarking. His often quoted definition of benchmarking is "the search for industry practices which will lead to exceptional performance through the implementation of these best practices" (Anand & Kodali, 2008, p. 258). Camp (1989) emphasises that benchmarking is a continuous, proactive and systematic process as functions were forced to not only understand the current internal world but also constantly assess the external one. This process is time- and labour-intensive (Spendolini, 1992), and in order to be successful requires major investments: Information need to be acquired with a suitable data collection methods, management support has to be created, and for the implementation the best practice has to be communicated across the whole organisation (Anand & Kodali, 2008).

Camp's (1989) definition lacks on indicating the range of applications for benchmarking, although existing literature shows that benchmarking can be conducted in different forms and on many different subjects. Main distinguishing factor is whether benchmarking is conducted with internal or external partners. External benchmarking can be with competitors, related or unrelated industries, whereas internal benchmarking stays within the organisation (Anand & Kodali, 2008; Drew, 1997). External benchmarking can help to find the global best practice in contrast to internal benchmarking, which risks falling short on it due to the limited view. However, external benchmarking is harder to achieve as information access to external industries and companies is often limited. Even if the information is available, adaption and implementation still might be a challenge due to culture differences (Drew, 1997; Southard & Parente, 2007). Internal benchmarking is only suitable if processes, services or products differ within an organisation, but still can be shared and adapted between organisational functions or departments (Southard & Parente, 2007; Spendolini, 1992). Internal benchmarking enables a better understanding of the organisation and its interrelations, which is the baseline for external benchmarking (Spendolini, 1992). To determine if external or internal benchmarking is suitable we can follow the benchmarking flowchart of Southard and Parente (2007) (see Appendix VI).

In addition to the benchmarking partners it is also critical to define the benchmarking subject. There are many subjects that can be benchmarked, but the main ones are product, process, function, strategy or performance benchmarking (Anand & Kodali, 2008; Anderson & McAdam, 2004; Dattakumar & Jagadeesh, 2003). Regarding to existent literature, we can see that benchmarking is successfully applied to improve operational performance (Binder, Clegg, & Egel-Hess, 2006; Chan, Henry, & Ralph, 2009; Voss et al., 1997), which is critical for our project. Binder(2006) showed how internal benchmarking can be implemented for the "packing and filling" process within the international chemical company BASF. Critical for applying benchmarking in practice, which is often not included in the theoretical models, is good communication and correct implementation of a feedback loop through all process steps. Biggest challenge of the implementation is, according to

Binder et al.(2006) , to compare and create similar entities as this is restricted by many factors in the "real-world".

One way of handling the complexity of "real world" is shown by Chan et al. (2009). They present a benchmarking framework which allows to incorporate many factors based on an AHP methodology for logistic performance of the postal industry. The AHP methodology enables to gain an overall performance score based on the relative weight of different criteria and sub-criteria, which can be both qualitative and quantitative. For benchmarking it is critical to gain support from different departments. The AHP methodology is a flexible tool which can incorporate new criteria. Also the relative weight can easily be changed according to wishes of stakeholders or the current operational strategy (Chan et al., 2009; Subramanian & Ramanathan, 2012).

Overall, benchmarking is a complex process, which however can be applied to different fields on different subject, enabling organisational learning and improvement. Anand and Kodali (2008) successfully incorporate those elements into their definition of benchmarking, which is based on a literature review of 35 academic articles on benchmarking. Therefore, we use their definition of benchmarking for our project with the focus on service and performance analysis:

*"a **continuous** analysis of strategies, functions, processes, products or services, performances, etc. compared **within or between** best-in-class organisations by obtaining information through appropriate **data collection** method, with the intention of **assessing** an organisation's current standards and thereby carry out **self-improvement** by implementing changes to scale or **exceed** those standards"* (Anand & Kodali, 2008, p. 259).

After having understood the concept of benchmarking, we determine the steps of a benchmarking process for our project in the following section.

The Benchmarking Process

The main elements of the benchmarking process are planning and defining the benchmarking elements, data collection and analysis as well as communicating and implementing best practices. While most benchmarking models contain those phases, we can see different numbers of steps and thus different degrees of complexity (Anand & Kodali, 2008). The pioneer of benchmarking frameworks is the ten step process presented by Camp (1989), which involves following steps:

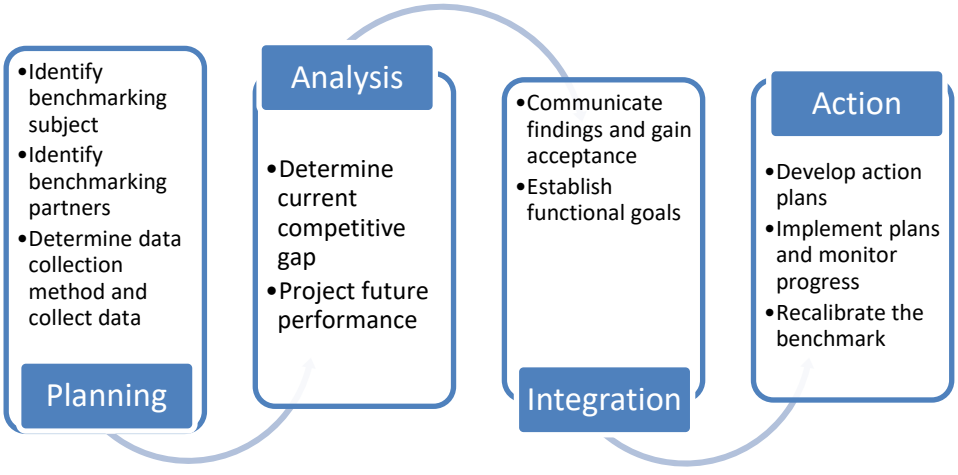


FIGURE 4.1: BENCHMARKING PROCESS (CAMP, 1989)

Camp's model is quite general and misses critical elements for a successful benchmarking implementation. For instance, Spendolini (1992) incorporates in the planning phase the step of forming a benchmarking team or Binder et al. (2006) present the model used by BASF, a large

multinational chemical company, for their internal benchmarking, established a feedback loop by adding between each main phase the step of communication.

Based on a benchmarking of benchmarking models, Anand and Kodali (2008) determined 54 best steps, which they divided into 12 phases (see Figure 4.1). We can see that this model is much more practical orientated by using project planning steps like team formation and action plans or involving more stakeholders into the steps "customer" or "management validation". Furthermore, the 54 steps are quite elaborated and provide exact tasks that one has to follow. However, this model is for general benchmarking and only based on theory. To ensure adaptability of this model into practice we compare it to the internal benchmarking process presented by Binder et al. (2006), who directly applied it to practice and, similarly to our project, included a clustering. Based on the comparison we define steps that we follow for developing the benchmarking model of mail delivery process.

The first two steps are similar to each other, but the sequence is not. Binder et al. (2006) argue that before selecting your resources, which includes the benchmarking team, one needed to know the aim and the goal of the project in order to select the right persons. By firstly determining the benchmarking subject and aim, we can define which expertise the benchmarking project team requires. Furthermore, instead of only establishing a team, which is suggested by Anand and Kodali (2008), we rather follow Binder et al. (2006) who establishes project management as it does not only incline to form a team and to divide the responsibilities, but also to define a project plan and detailed milestones. Thus overall, we will firstly define subject, aim and goal of the project and subsequently establish the project management.

The third step is to determine the benchmarking focus and benchmarking partners. To ensure value creation of benchmarking, one should set the focus based on the business as well as on the customer needs. Binder et al. (2006) select the benchmarking focus sorely based on the impact on customer satisfaction. However, Anand's and Kodali's (2008) model emphasises to firstly identify the users of the benchmarking information and subsequently select the focus based on the needs of all stakeholders (users, the business and the customers). Identifying the needs of users instead of only the ones of customers ensures that findings of the benchmarking are useful and value adding.
After defining the benchmarking focus, we can select the benchmarking partners, thus the people who compare their performance. At an internal benchmarking partner selection is quite limited and hence one can easily select partners based on the subject and aim of the benchmarking (Binder et al., 2006). However, to identify potential partners for an external benchmarking, one first needs to gain knowledge on the external environment. We can see this in the benchmarking process model of Anand and Kodali (2008) as well, where the partner selection is later on in the process after analysing a self-analysis with a competitive positioning.

The fourth step for our benchmark model is to drive critical success factors (CSF) of the benchmarking subject, which we do based on a top-down and bottom-up approach. CSFs "are the limited number of areas in which results, if they are satisfactory, will ensure successful competitive performance for the organisation" (Rockert, 1979, p. 85). Therefore, it is critical to continually measure the performance within that area and make that information available within the company (Rockert, 1979). Anand and Kodali (2008) suggest to identify CSF "based on the subject of benchmarking, strategic intent, core competencies and capability" (p. 283). However, it is not only important to determine CSF from a strategic level, but also to incorporate the operational level by interviewing employees directly involved in that process (Binder et al., 2006; Korpela & Tuominen, 1996). Overall, by combining Blinder et al. (2006) and Anand and Kodali (2008) we gain a top-down and bottom-up approach, which is also suggested by Bourne and Neely (2003) for designing a performance measurement system.

The next step, Step 6, is to understand the current situation in order to determine subsequently drivers behind the CFS. Therefore, we need to analyse existing information on that benchmarking subject (Anand & Kodali, 2008). Binder et al. (2006) suggest to create a flowchart to understand the

process as well as to visualise the process in order to illustrate the interdependence between structure and process.

After having understood the situation, we can determine performance measures and their metrics that are behind the CFS in Step 7. The term performance metrics refer to specifying the performance measure in terms of the measurement unit, the intervals of measurement and the formula for its calculation (Anand & Kodali, 2008; Binder et al., 2006; Neely et al., 1995). The exact elements of the performance metrics are defined in Section 4.2.
Anand's and Kodali's process model (2008) suggests to directly measure the performance in order to understand and assess the inherent capabilities. Only by understanding your own capabilities, you can compare it with others (Spendolini, 1992). However, in reality information acquisition on your own company is often a challenge and therefore defining and measuring the performance measures should be in a separate step, which is the case in Binder et al. (2006).

Before measuring the performance, we need to define the clustering. Binder et al. (2006) determined their clusters in an earlier stage of the benchmarking process without knowing the exact performance measures. Based on different properties suggested by stakeholders of the process and on-site visits, they defined primary and secondary properties which they used for defining the clusters. However, the aim of clustering is to minimize the variance within a cluster (Tan et al., 2005b). To do so, one needs to know which factors or properties lead to the variance before clustering them. Therefore, we conduct the clustering in Step 8, after having defined performance measures.

Step 9 concerns data collection and verification. Both processes, the internal as well as the general one, require to define the information need and to determine a suitable data collection method. A difference however is that in contrast to external benchmarking, internal benchmarking contains lower barriers for data collection. In the general benchmarking process, we can see many steps concerning legal formalities including setting a reciprocal and non-disclosure agreement (Anand & Kodali, 2008), which slow down the process and are not necessary for our purpose.

The last steps are for both benchmarking processes identical: analysing and comparing the performance, defining best practices and performance gaps, developing an action plan, implementing the defined actions and finally monitoring the progress and continuously improving by motivating the benchmarking partners through a structure rewards system and recalibrating the benchmarking model.
An overview of the comparison and our final benchmarking process is given in Table 4.1. The sequence of the steps is indicated by their numbering.

TABLE 4.1: BENCHMARKING MODEL PROCESS

Anand & Kodali 2008 General benchmarking	Binder et al. 2006 Internal benchmarking	Final steps selected for this project
1. Team formation	2. Establish project management	1. Subject identification
2. Subject identification	1. Set aims and goals of the project	2. Establish project management
3. Customer validation (stakeholder analysis, validate benchmarking topic)	3. Define specific focus, partners and type of benchmarking	3. Define specific focus and benchmarking partners
4. Management validation (benchmarking proposal)		X *(management validation after each step)*
	5. Develop clusters of filling lines	X *(clustering depends on performance measures thus later on in this process)*
5. Self-analysis (current situation, critical success factors)	4. Identify critical success factors of the process	4. Identify critical success factors of the process
	6. Understand the current situation	5. Understand the current situation
	7. Develop performance metrics	6. Develop performance metrics
6. Partner selection	3. Define specific focus, partners and type of benchmarking	7. Define the clusters
7. Pre-benchmarking activities (contact partner, data collection method and agreement with partners)	8. Conduct interviews and collect data	8. Data collection and verification (method)
8. Benchmarking	9. Analyse and compare	9. Analyse and compare
9. Gap analysis	10. Define best practices and performance gaps	10. Define best practices and performance gaps
10. Action plans	11. Develop and evaluate improvements	11. Action plan
	12. Communicate results	
	13. Develop roadmap for detailed implementation	
11. Implementation	14. Implement changes	12. Implementation
12. Continuous improvement	15. Monitor success of implementation	13. Monitor and continuous improvement

4.2. Performance Measurement and Performance Measures

In this section, we will provide a definition and criteria for a good performance measure and elements of the performance metrics to be able to execute Step 6 of the benchmarking process.

Performance measurement is the process of quantifying an action; (Neely et al., 1995). To do so we need to set certain performance measures, which can be defined as "as a metric used to quantify the efficiency and/or effectiveness of an action" (Neely et al., 1995, p. 1229).

There are various performance measures. One characteristic which can differ is the level of detail: More detailed performance measures, like order lead time, have a high diagnostic power. They enable a more descriptive and complete view that clarifies the gap between planning and realisation. However, detailed measurements require a close monitoring and a highly frequent measuring. Generic operations performance measures, in contrast, like cost, quality, or speed, enable a broad view on the performance and show more the strategic relevance (Slack, Chambers, & Johnston, 2010). For benchmarking a combination of both would be suitable: first a more generic operational performance measure to illustrate their overall performance and then zooming in on it by using detailed performance measures of that field to identify performance gaps or best practices.

Not only can the level of measurement vary, but also the scope. Performance measures were used originally in the financial area; however, this encourages short termism and does not consider the external environment. Later on the focus shifted towards quality and consumer satisfaction. But as the environment is getting more dynamic, companies should also consider performance measures on long-term value creation, thus on innovation and knowledge management (Anderson & McAdam, 2004).
To define the scope of performance measurement different performance measuring frameworks are presented. There are several frameworks within the academic literature to categorise performance measures (Bourne & Neely, 2003). Some focus on the different levels of performance measures (e.g. performance pyramid by Lunch & Cross (1991)), others on the different organisational perspectives (e.g. balanced score card by Kaplan & Norton (1992), performance measurement matrix by Keegan, Eiler & Jones (1989)) and some on the horizontal (information or material) flow within an organisation (e.g. Inputs, processes, outputs, outcomes by Brown (1996)). Our aim is not to select performance measures that reflect the whole organisational performance, but to focus solely on the mail delivery process from a process manager perspective. Frameworks like the balanced score card presented by Kaplan & Norton (1992) or the business excellence framework are helping to translate the overall corporate strategy into a balanced set of performance measures, however those frameworks are too broad for our purpose. In contrast, the framework of Brown (1996) is very specific and focusses on processes. He emphasizes the difference between input, processing system, outputs, outcome and goal by suggesting that each step had its own specific performance measures (Neely et al., 2000). By applying this framework to the mail delivery process, we can gain clarity on internal factors; however this neglects external factors like customer satisfaction which are quite relevant for the process managers.

Slack et al. suggest a framework which incorporates different perspectives, but still keeps a focus on the operational level (see Figure 4.2). Using this model enables us to combine all three procedures suggested by Bourne and Neely (2003) for developing performance measures, the needs led procedure which considers the requirements of customers, business and stakeholders which we can see within the top-down and market requirement perspective, and the audit led procedure which follows the bottom-up perspective. Finally, within each perspective we can apply theoretical models which would be the model led procedure. For instance, to analyse the day-to-day experience we can apply the framework for processes of Brown (1996).

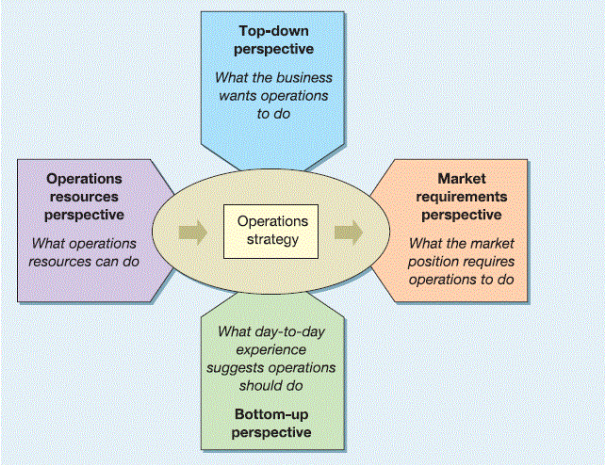


FIGURE 4.2: THE FOUR PERSPECTIVES ON OPERATIONS STRATEGY (SLACK ET AL., 2010)

Overall, we use the four perspectives on operations strategy framework as a guideline to derive performance measures. However, before being able to define performance measures we have to clarify some criteria for a good performance measure as well as the elements of the metric, which we do in the following section.

Criteria

To ensure effective performance measures we use the 22 recommendation given by Neely et al. (1997). They selected based on the extensive academic literature review on performance measurement of Neely, Gregory and Platts (Neely et al., 1995) ten different papers and books to develop those recommendations. We summarised the most distinctive recommendations below and categorised them into five different areas: the development, aim, calculation, presentation and implementation of performance measures.

Performance measures should …

1. Development:
    a. …be derived from strategy and reflect the business process – i.e. both the supplier and customer should be involved in the definition of the measure.
2. Aim:
    a. …have an explicit purpose and relate to specific goals (targets)
    b. …focus on improvement
    c. … provide information
3. Calculation
    a. …be clearly defined and be based on an explicitly defined formula and source of data
    b. …employ ratios rather than absolute numbers and be based on trends rather than snapshots
    c. …be based on quantities that can be influenced, or controlled, by the user alone or in co-operation with others
    d. …use data which are automatically collected as part of a process whenever possible
4. Presentation:
    a. … be simple to understand and should be reported in a simple consistent format
5. Implementation:
    a. …be part of a closed management loop
    b. …provide timely and accurate feedback and should be objective – not based on opinion
    c. ..be consistent

(Neely et al., 1997, p. 1337)

The aim of Neely et al. (1997) is that the recommendations can be applied universally, however for our project we need to make some specifications and adaption of the Recommendation 1a, 2a, 5a, 3b and 3d:

Recommendation 1a: Performance measures should be derived from strategy, but as we do not define performance measures for the corporate level, we should focus on the strategy for the mail delivery process. This strategy is defined by the logistic strategy department based on the corporate strategy and the requirements of the mail delivery process.

Recommendation 1a: Scope of our research is to define performance measures for the mail delivery process which starts at the HUB or depot and ends at the customer. Thus including suppliers would be out of our scope. However, we consider the customers as for service provider, which the mail delivery department is, customer satisfaction is critical (Chan et al., 2009; Grönroos & Ojasalo, 2004).

Recommendation 2a: Benchmarking is more focused on learning and exchanging (Dattakumar & Jagadeesh, 2003; Voss et al., 1997; Yasin, 2002) than on fulfilling specific targets. Achieving certain targets is rather important in MJ dashboards at PostNL, which compares budgets to the realisation. To motivate process managers, we should provide a certain goal. In benchmarking the goal is to adapt best practices from others. Thus each goal of a performance measure is based on the best score within the benchmarking and given per cluster.

Recommendation 5a, 3d: Those recommendations are highly important and we do consider them, however as we are not implementing the benchmarking model, but only design a prototype, designing an information system that could automatically retrieve the information is out of our scope. Furthermore, we do make recommendations for the management of the benchmarking, but do not design and implement a specific management loop.

Recommendation 3b: Evaluating trends rather than snapshots gives the advantage to assess and to compare performance improvements between the different benchmarking partners, however snapshots can give insights into one moment and enable to zoom in and determine the root of the problem. By combing snapshots of different moments, we are still able to show the performance trends.

Elements of the Performance Metrics

Based on the comparison of the two benchmarking processes, we should define as part of the benchmarking process a metric for each performance measure to ensure that it is sufficiently specified. Binder et al. uses four elements for the performance metrics: related CSF, title, unit of measurement and formula of the performance measure. Anand and Kodali (2008) add one more element: the interval of measuring the performance. Still, those elements are not sufficient as the recommendation by specifying only those elements are not fully satisfied. For instance, those elements do not clarify the relevance (Recommendation 2a) of the performance measure.

Neely et al. (1997) developed a performance measure record sheet, which incorporates ten elements derived based on the 22 recommendations. This framework (see Table 4.2) incorporates not only those recommendations, but also covers the elements named by Binder et al. (2006) and Anand and Kodali (2008). Using this framework for our performance metrics ensures that performance measures are sufficiently specified, which is a critical part of Step 6 in our benchmarking process.

TABLE 4.2: PERFORMANCE MEASURE RECORD SHEET (NEELY ET AL., 1997)

Element	Related recommendation	Description
Title	3a, 4a	The title should not only specify what has to be measured but also why. It should be easy to understand and self-explaining.
Purpose	2a	To clarify the relevance of the performance measure one should briefly state the rational reasoning behind it.
Relates to	1a, 2a, 2b	The performance measure is related to the business (or in our case the mail delivery department) objective to show the value and the connection with the overall context.
Target	2a, 2b, 2c, 3c, 5a	The performance measure should provide a certain target, which we base on the best performance within the benchmarking.
Formula	1a, 3a, 3b, 3c, 4a, 5b	The formula enables to calculate the performance measure but also determines the unit of measurement.
Frequency	2c, 4a, 5b, 5c	Includes the frequency of measure and of review
Who measures?	3c, 3d	Specifies the person that collects and reports data
Source of data	3a, 3b, 3d, 4a	Clarifies the source of the raw data to ensure consistency in the measurement over time
Who acts on the data?	2a, 2c, 3c	Names the person who is responsible for acting on the measure
What do they do?	2a, 2c, 3c	Specifies actions that could influence the outcome of the performance measure
Notes and comments		

4.3. Clustering

In this section we describe critical elements of a cluster analysis and discuss existing academic literature in that field. Goal is to determine methods that can be applied to our clustering for the benchmarking model and by that answering Research Question 3e: *What solution approaches for clustering exist in the literature?*

Therefore, we first define cluster analysis in Section 4.3.1. and describe the basic idea and concept. Subsequently, we investigate approaches for selecting and weighting cluster attributes in Section 4.3.2. After this, we analyse current clustering techniques of the academic literature. By considering and comparing their advantages and drawbacks we assess their usefulness for our problem context. Finally, we investigate existing literature on cluster validation methods which can be applied later on for evaluating our clustering for the benchmarking model.

4.3.1. Cluster Analysis

Cluster analysis is part of data mining. Similar to other basic techniques of data mining, which are classification, visualisation, summarization and prediction, the aim is to extract relevant features from data and to find useful patterns in the feature space (Tan et al., 2005b). In particular, cluster analysis, often referred to as clustering, is the study of methods and algorithms that identify meaningful and useful groups (clusters) in a dataset based on information describing the object and its relations (Berkhin, 2006; Halkidi, 2001; Jain, Murty, & Fylnn, 2000; Tan et al., 2005b; Xu & Wunsch, 2005).

The overall goal of clustering is to create highly distinctive clusters by achieving high similarity within a cluster and a high difference between them. In contrast to classification, which is also a technique concerning grouping objects, where models for grouping are developed based on objects with already known class labels, there is no antecedent information on classes within clustering, but clustering sorely relies on information describing the objective. Cluster algorithms have to detect data patterns on their own, which is often referred to as unsupervised learning, and are thus explanatory of nature.

Cluster analysis is a widely used concept with a rich history. It has been introduced in fields like psychology, biology or marketing in the mid-20th century and is still gaining importance especially in fields like information retrieval, image processing and pattern recognition. In particular, through rise of data creation and availability or “big data” and increasing requirements for data mining, the need for clustering in the last year increased in multiple domains (Halkidi, 2001). It is notable that the general clustering techniques introduced mid and end-20th century are still used in recent academic literature as a fundament of more advanced algorithms (Jain, Murty, et al., 2000). Thus, in the following literature review on clustering techniques in Section 4.3.4. we will first discuss the traditional algorithms and subsequently the recent advances.

Critical for the output of the cluster algorithm are the input parameters. One input that is required by all cluster algorithms are the objectives and their attributes. However, as we do not have any cluster descriptions or labels, we have to decide on our own on the input attributes for comparing the similarity between the objectives. In clustering, objects are mostly described by a multidimensional vector, where each dimension represents one attribute. Within cluster analysis there is a distinction made between high and low dimensional data as they require different approaches. Usually, one strives for low dimensional data, not only to keep (computational) complexity low, but also because traditional similarity measures, which are often based on distance measures like the Euclidean distance, are not meaningful anymore in high dimensional data. In order to keep dimensions as low as possible, it is critical to select only relevant and representative attributes (Jain, Murty, et al., 2000; Jing, Ng, & Huang, 2007; Kriegel, Kröger, & Zimek, 2009; Xu & Wunsch, 2005). Depending on the area, high dimensional clustering might be important, for instance gene expression or text documents analysis (Kriegel et al., 2009), however, for our problem context we assume that this is not relevant and thus take no special focus on high dimensional data clustering. Still it is important to assess if attributes are relevant for the clustering, which we will discuss in Section 4.3.2, followed by an overview on distance measures used to indicate the similarity in Section 4.3.3.

As no antecedent knowledge on existing clusters and their attribute is given, which could confirm our results of cluster analysis, it is important to critically evaluate the outcome (Berkhin, 2006; Halkidi, 2001; Xu & Wunsch, 2005). The academic literature presents different evaluation and validation techniques which we will discuss in Section 4.3.4.

Overall, those critical elements of cluster analysis – attribute selection and cluster validation – form steps of the clustering process. Even though clustering is used in different fields, there is a general agreement within the academic literature on the main steps: firstly, selecting relevant attributes and pre-processing the data, secondly, determining an adequate clustering technique and algorithm – a

clustering technique is a general strategy for solving a clustering problem, whereas an algorithm is a specific instance within that strategy (Jain & Dubes, 1988) –, thirdly, validating the outcome of the clustering and finally interpreting the results (Halkidi, 2001; Jain & Dubes, 1988; Xu & Wunsch, 2005). Jain and Dudes (1988) emphasise that in cluster analysis all steps are equally important and thus should be followed carefully. Xu and Wunsch (2005) summarised those steps in Figure 4.3, which we will use as a guideline for determining the clusters for the benchmarking model.

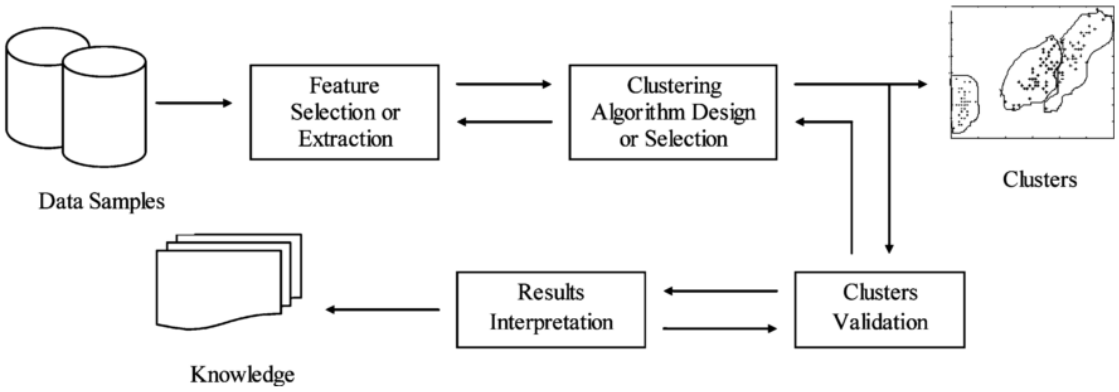


FIGURE 4.3: CLUSTERING PROCESS (XU & WUNSCH, 2005)

In the following we will discuss solution approaches of the academic literature for the first three steps. But before doing so we will clarify the terminologies within cluster analysis by defining terms that we use in our paper and providing the synonyms, which are often used, in brackets (Berkhin, 2006): A given dataset X contains N objects (data points, instances, patterns, cases), where each object consists of a vector of d measurements $x_i = (x_{i1}, x_{i2} \dots, x_{nd})$, where x_{ij} refers to the measurement of attribute j (feature, variable, dimension, component, factor) of the i^{th} object within a d dimensional objective space. Objectives and their attributes are summarised in a n×d objective matrix. Clustering creates a set of K clusters, $C = \{C_k, k = 1, \dots, K\}$ and assigns a class label l_i to each object x_i with $l_i \in \{1, \dots, K\}$, where K is the number of clusters. A n×n proximity matrix quantifies the similarity of object, with $object_j$.

4.3.2. Cluster Attributes (Selection, Weighting)

The first step of cluster analysis is to select and/ or to extract attributes that we want to incorporate in the clustering. Attribute selection refers to identifying useful attributes from a given set of candidate attributes, while extraction means to create novel and silent attributes by transforming the given attribute set (Jain, Murty, et al., 2000).

Attribute extraction is especially important if we want to reduce the number of attributes and to extract only important components that contribute to the cluster division. Conducting a feature extraction always causes loss of information and hence incorporates the risk of distorting the real clusters (Xu & Wunsch, 2005). Therefore, it should only be applied if it is really necessary. One of the most common techniques used is principle component analysis which extracts the most important components for describing the variance of data (Hall & Holmes, 2003; Jain, Murty, et al., 2000; Raftery & Dean, 2006; Xu & Wunsch, 2005). Some use the principal component technique also to reduce the data to two or three dimensions and to conduct a clustering based on a visualisation. The disadvantage is that such an extreme process comes with many pitfalls (Kettenring, 2006). Chang (1983) emphasises that principal components might account for the most variability, but do not necessary contribute to a clear cluster structure (Raftery & Dean, 2006). Furthermore, the new reduced dimensions might be harder to interpret and to gain an understanding of the clusters in their original space (Jing et al., 2007). Therefore, data extraction is often only applied to high dimensional cases. An overview of attribute extraction techniques is given in Xu and Wunsch (2005) or Kriegel et al. (2009).

Attribute selecting techniques proposed within the clustering literature are mainly focused on predictive rather than descriptive mining and thus considers in particular to what extent certain attributes are necessary to define the cluster structure (Berkhin, 2006). If attributes provide no new information for the cluster structure, it not only leads to higher complexity, but also might mask the true cluster structure, which makes it harder for an algorithm to discover the true clusters. Therefore, often authors refer to them as "masking" attributes (Brusco & Cradit, 2001; Steinley & Brusco, 2008a). A possible continuation of attribute selection is not only to eliminate those masking attributes, but also to give attributes weights based on their ability for discriminating the clusters. However, as Jain (2010) emphasises to always keep in mind the application need for the clustering, it might not even be required to create highly distinctive clusters. Thus overall, for a good attribute selection one should not only evaluate the relevance of the attribute to describe an object, but also the extent to which it contributes to the cluster structure.

Overall attributes should fulfil the following criteria:

1. **relevance for the problem context**
   Only relevant and representative attributes should be selected (Jain, Murty, et al., 2000; Jing et al., 2007; Kriegel et al., 2009; Xu & Wunsch, 2005). Therefore, a first selection of potential attributes can be made by understanding and analysing the problem context to limit and subsequently select only those attributes that are influencing the problem context. (Huang, Ng, Rong, & Li, 2005) In other words, we should asses which attributes might characterise the objects in a certain problem context most (Xu & Wunsch, 2005).
2. **distribution of the data**
   The aim of clustering is to define distinct groups within the data. Uniform distributed attributes do not provide useful information on the cluster structure and most likely result in a spurious cluster (Law, Figueiredo, & Jain, 2004; Tibshirani, Walther, & Hastie, 2001). In order to evaluate if a distribution of an attribute helps to determine the cluster structure, Steinley and Brusco (2008b) presents a relative clusterability index (CI), which is built on the variance-to-range ratio of the attribute (j):

$$CI_j = \frac{12 * var(x_j)}{(r(x_j))^2} \quad (1)$$

$$r(x_j) = \max(x_j) - \min(x_j) \quad (2)$$

   For a uniform distribution CI would be one as the variance is equal to $\frac{1}{12} * (\max(x_j) - \min(x_j))^2$. The more x_j goes beyond an uniform distribution, the higher the clusterability index gets(Steinley & Brusco, 2008a). In order to compare the clusterability among different attributes a relative clusterability is defined:

$$RCI_j = \frac{CI_j}{\min_j(CI_j)} \quad (3)$$

   The attribute which is least clusterable gains a RCI of 1, all RCIs of the remaining attributes indicate how many times more, relative to the least clusterable attribute, they are clusterable.
   While a low CI indicates that the attribute is likely to provide not much information on the cluster structure, a high CI does not always represent a strong cluster attribute, because they might not work well in combination with other attributes (Steinley & Brusco, 2008a). This can be supported by Law et al. (2004), who emphasise that even if an attribute had a high variance, it does not automatically mean that it would contribute to more distinctive clusters as the variance might be independent of the intrinsic grouping of the data. In Figure 4.4 (Law et al., 2004, p. 1155), we can see that although x_1 has a higher variance than x_2, it does not improve the clustering but rather blur the cluster structure and can rather be seen as a "masking" attribute (Brusco & Cradit, 2001; Steinley & Brusco, 2008a).
   Therefore, next to the clusterability index we have to control, if the variance goes along with the intrinsic group or, in other words, if a cluster tendency exists within the data.

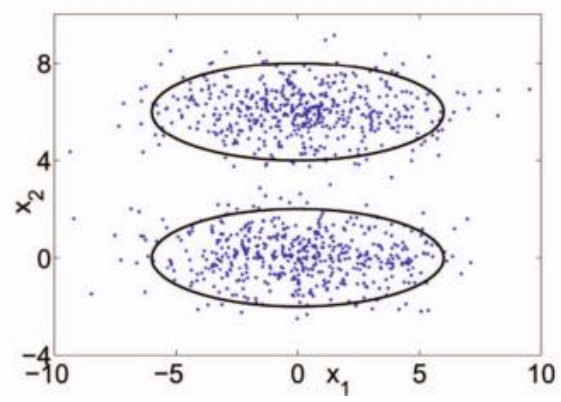


FIGURE 4.4: VARIANCE AND CLUSTER IDENTIFICATION (LAW ET AL., 2004, P. 1155)

3. **cluster tendency**

Clusters can only be found in a data set, if a non-random structure actually exists or, in other words, if there is a cluster tendency within the data. It is critical to assess the clustering tendency as cluster algorithms can even identify clusters in data which do not possess natural clusters (see Figure 4.5) (Jain, Murty, et al., 2000; Tan et al., 2005b).

The exact cluster tendency within the data depends on the distribution and the correlation of the data set which again depends on the chosen cluster attributes and their correlation (Halkidi, 2001). Evaluating the cluster tendency is often an interactive process by adding and removing attributes and re-evaluating the data pattern. To discover if an attribute contributes to the cluster structure, an efficient and simple way is to plot the data points in a scatterplot and estimate based on the pattern if attributes contribute to a clear cluster structure. Such a scatter plot can also help estimate the number of clusters within the data. However, this is limited to two and three dimensional data. Therefore, attribute selection is often a trial-and-error process which evaluates and tests various subsets of attributes.

A first indication for correlation and cluster tendencies can be gained by clarifying the pairwise relation between attributes. Firstly, one should assess if a linear correlation exist between attributes as this can distort the distance measure between objects (Berkhin, 2006; Jain, Murty, et al., 2000). For cluster analysis one strives to minimise the multicollinearity in particular if all the attributes should have equal weight. One approach is the principal component analysis, which however, as mentioned above, has many pitfalls. Another approach is to use the Mahalanobis distance which assigns different weights to attributes based on the variance and pairwise linear correlations (see Section 4.3.3) (Jain, Murty, et al., 2000; Ketchen & Shook, 1996). Finally, one should create scatter plots of all possible pairs of attributes as this can already provide an indication of the relevance of a variable and its potential contribution to the clustering (Raftery & Dean, 2006).

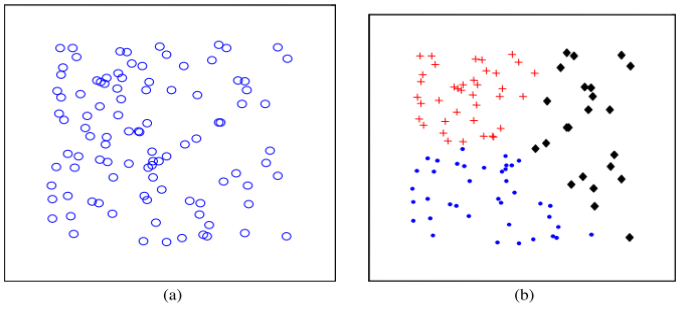


FIGURE 4.5: CLUSTERING TENDENCY A) DATA SET WITH NO NATURAL CLUSTERS B) CLUSTERS IDENTIFIED BASED ON K-MEANS (K=3)(JAIN, 2010, P. 658)

4. **easy to extract and interpret**
   In order to conduct a clustering we require the data on all attributes and for each cluster object. In practice, not all information is available and/or too expensive to acquire. Therefore, we should only select those attributes where the information is easy to extract or within reasonable expenditure obtainable. Furthermore, as the goal of clustering is to find meaningful groups within the data, the outcome should be understandable for the end user, which requires among other aspects that the (combination) of attributes are easy to interpret (Xu & Wunsch, 2005).

There are different approaches that we can follow to select and to weigh the variables which consider especially Criteria 2 and 3. Within the clustering literature weighting and selecting attributes are often combined as attribute weighting often indicates the usefulness of an attribute and helps to decide if it should be select or not (Kettenring, 2006; Steinley & Brusco, 2008a). We can distinguish between model based and non-model based selection and weighting techniques. Steinley's and Brusco's (2008b) extensive comparison of variable selection techniques shows that model based approach performs significant worse than the none-model based approach. Therefore, we focus on the non-model based approach.

There are several variable selection approaches and also several studies comparing them (see (Gnanadesikan, Kettenring, & Tsao, 1995; Raftery & Dean, 2006; Steinley & Brusco, 2008b). However, being able to conduct an approach depends on the degree of information and the ability to give accurate estimations on clustering (Kettenring, 2006; Raftery & Dean, 2006). For each level of information we selected the most efficient weighting approach, which we summarised in Figure 4.6.

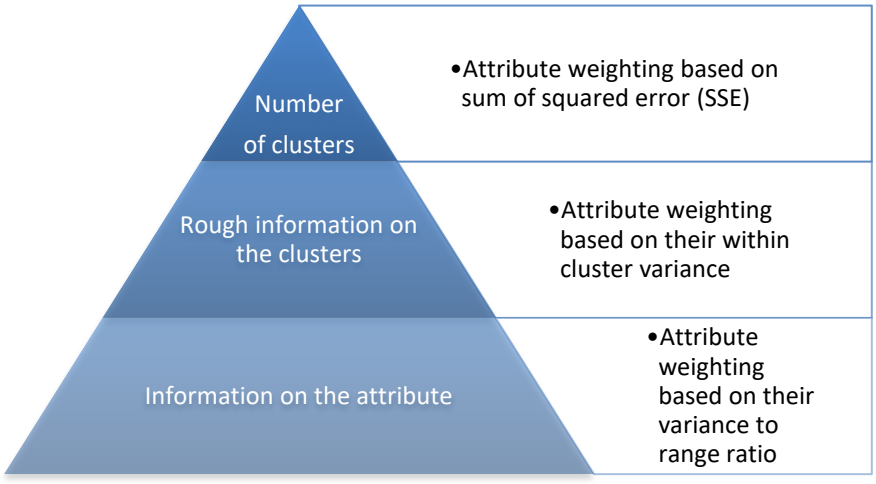


FIGURE 4.6: LEVEL OF INFORMATION AND VARIABLES WEIGHTING APPROACH

If we have only the attributes without being able to give any estimation on the grouping or distribution many attribute weighting approaches cannot be used. One approach which still can be applied is the variance-to-range ratio weighting presented by Steinley and Brusco (2008a) which is based on the (relative) clusterability index of the attribute:

1. calculate for each attribute (j) the relative clusterability index (RCI_j) (see Formula 3 above)
2. Transform x_j with the traditional z-score method to z_j^1 so that each attribute has the same scale and a variance of one:
$$z_j^1 = \frac{x_j - sample\ mean}{sample\ standard\ deviation} \quad (4)$$
   determining the new range for each attribute, called $r(z_j)$
3. Reweight the attribute of z_j^1 with weight w_j such that the RCI_j holds, which is computed by:
$$z_j^2 = z_j^1 w_j \quad (5)$$

$$w_j = \sqrt{\frac{RCI_j[r(z^1_{\min})]^2}{[\mathrm{r}(z^1_j)]^2}} \quad (6)$$

Where $z^1_{\min}$ is the z-score transformation corresponding to CI_{min}

An empirical comparison, conducted by Steinley and Brusco (2008b), showed that this weighting technique helps to improve the cluster structure when skewed random noise is present. Moreover, weighting on standard deviation and range can give an indication on the quality of an attribute, but has to be used with caution as like mentioned above it does not automatically mean that it will contribute to a more distinctive clusters (Law et al., 2004). This can be supported by Gnanadesikan and Kettenring (1995) who compared nine weighting and selection methods and conclude that weighting sorely based on standard deviation or range has many shortcomings, but was still better than autoscaling. Their study showed that weighting methods based on carefully estimated within-cluster and between-cluster variability are more efficient and also better than equal weighting. Therefore, no information on the exact cluster structure is needed; however one has to be able to determine "likely" within-cluster pairs. The weighting matrix (attribute x weight of attribute) provides higher weights to attributes that have expected low within-cluster variance. For the exact way of calculation we refer to Gnanadesikan and Kettenring (1995).

If the number of clusters can be estimated, we can apply the k-means algorithm (outlined in Section 4.3.4) and conduct a sort of sensitivity analyse by keeping everything constant except the set of attributes and compare the cluster outcome based on certain criteria (Halkidi, 2001). Steinley and Brusco (2008b) compare eight different selection methods. The best three approaches follow such a sensitivity analysis. The best one is the attribute selection based on the variance-to-range ratio (Steinley & Brusco, 2007), which is proposed by the same authors who also conducted the comparison. The third and second best approach (HINoV by Carmone et al., 1999 and the VS-KM by Brucsco&Cradit, 2001) perform nearly as good, however are more sensitive to skewed data. Another advantage of the attribute selection method of Steinley and Brusco (2007) is that it combines Criteria 2 and 3 by pre-screening the attributes on their clusterability index before starting the actual selection process. As for our benchmarking we do not know the exact number of clusters, we will not outline this algorithm, but refer to Steinley and Brusco (2007) for a detailed outline of the method.

Overall, we can see that the more information we have and the better estimation we can make, the more precise, but also the more complex the attribute weighting and selection approach gets.

4.3.3. Distance Measures

The method of assessing similarity between objects used depends on the attribute type; (Jain, Duin, & Mao, 2000; Xu & Wunsch, 2005) binary, nominal, ordinal and continuous. Binary variables are mostly measured based on a similarity measure ($S_{i,j}$) comparing the number of simultaneous absence (n_{00}) or presence (n_{11}) attributes in both objects, object j and object i, with the number of attributes only present in one of the objects (n_{01}, n_{10}):

$$S_{i,j} = \frac{n_{11}+y*n_{00}}{n_{11}+y*n_{00}+w(n_{01}+n_{10})} \text{ with y = [0,1], w = (0,2)} \quad (7)$$

Depending on the priorities one can set different weights for each sort of match or mismatch by adjusting the values for parameter w and y. With the parameter w one can give dissimilarity between objects a different weight than similarity. With parameter y = 0, we can set the focus on co-occurrence rather than co-absence.

This measurement can also be applied to nominal data, as long as it has only two states. Otherwise, a matching criterion has to be defined for each variable. The similarity between objects is then based on the average number of matches between all attributes: (Xu & Wunsch, 2005)

$$S_{i,j} = \frac{1}{d}\sum_{l=1}^{d} S_{ijl} \;\; with \;\; S_{ijl} = \begin{cases} 0, if\ i\ and\ j\ do\ not\ match \\ 1, if\ i\ and\ j\ match \end{cases} \quad (8)$$

The most common technique used to assess the similarity is a distance function, which is limited to continuous attributes (Jain, Murty, et al., 2000; Xu & Wunsch, 2005). However, ordinary attributes are often adapted by assigning to each value based on a predefined standard and an order number so that a distance function can be used (Xu & Wunsch, 2005). There are various kinds of distance functions. Most common one within cluster analysis is the Euclidean distance as it is a simple and fast distance function. In the following table (Table 4.3) we will summarise frequently used distance measures for continuous attributes for low dimensional data. For more details we refer to Xu and Wunsch (2005) who also provide information on similarity and distance measures for document clustering or high dimensional data.

TABLE 4.3:DISTANCE MEASURES FOR CONTINUOUS ATTRIBUTES

Measures	Calculation	Characterisation	Available
Euclidean distance	$d_p(x_i, x_j) = (\sum_{k=1}^{d} \lvert x_{i,k} - x_{j,k} \rvert^p)^{\frac{1}{p}}$ with p = 2 (9)	special case of Minkowski distance with p=2, most common metric in cluster analysis, tends to form hyperspherical clusters, variables with a large range and variance gain more importance, assumes no linear correlation between attributes	SPSS, WEKA
Squared Euclidean distance (SED)	$d_p(x_i, x_j) = \sum_{k=1}^{d} (x_{i,k} - x_{j,k})^2$ (10)	same characteristics like Euclidean distance, except that here objects that are further apart gain more emphasis	SPSS
Manhattan/ City-block distance	$d_p(x_i, x_j) = \sum_{k=1}^{d} \lvert x_{i,k} - x_{j,k} \rvert$ (11)	special case of Minkowski distance with p=1, tends to form hyperrectangular clusters, attributes with a large range and variance gain more importance, assumes no linear correlation between attributes.	SPSS, WEKA
Mahalanobis	$D_{ij} = (x_i - x_j)^T S^{-1} (x_i - x_j)$ with S = within-group covariance matrix T = transpose of the matrix (12)	does not assume independency between attributes, tends to form hyperellipsoidal clusters, similar to Euclidean distance if attributes are not correlated, may rise computational burden	
Chebyshev/ Sub distance	$D_{ij} = \max_k(\lvert x_{i,k} - x_{j,k} \rvert)$ (13)	special case of Minkowski distance with p $\rightarrow \infty$, attributes with a large range and variance gain more importance, assumes that only the attribute with the largest difference is relevant, assumes attributes independency	SPSS, WEKA

Manhattan distance, Euclidean distance and Chebyshev are special sorts of Minkowski distance, with respectively p=1, p=2 and n $\rightarrow \infty$. Depending on the preferred cluster shape, we can select Manhattan (hyperrectangular) or Euclidean distance (hyperspherical). Furthermore, if we want to emphasise cluster objects which are further apart we should use the Squared Euclidean Distance. Finally, if it is only relevant to minimise the maximum difference between objects and their attributes, we should apply the Chebyshev distance as this function only returns the maximum absolute difference in coordinates. (Burnham & Anderson, 2004; Jain, Duin, et al., 2000)

There are two main disadvantages of the Minkowski metrics; firstly it assumes no linear correlation between attributes, and secondly it provides unequal weight to attributes if they differ on scale or variance. If collinearity exist within the data we can apply the Mahalanobis distance, "which is

equated to the Euclidean distance in a transformed whitened space" (Xu & Wunsch, 2005, p. 659). The Mahalanobis distance function normalises the attribute based on their covariance matrix and thus gives weight to the different attributes based on their variance and pairwise linear correlation. However, this function comes with high computational costs (Jain, 2010; Milligan & Cooper, 1988; Xu & Wunsch, 2005) and is also not available in statistical tools like SPSS or Weka.
The second disadvantage of Minkowski metrics, unequal emphasis of attributes, is often tackled by normalising attributes to a common range or variance (Jain, Duin, et al., 2000; Milligan & Cooper, 1988). However, as mentioned above, some intentionally overemphasize attributes which help to identify more distanced clusters (Kettenring, 2006). Milligan and Cooper (1988) compared eight different approaches with the conclusion that it is most efficient to divide the attribute by its range. They point out that standardizing attributes based on the traditional z-score formula (see Formula 4 in Section 4.3.2) was unadvisable. It leads to a common range and variance; however, as discussed above, variance can be critical to determine distinctive clusters. Thus, by this approach a high degree of cluster information will be lost. Of all eight approaches, dividing attributes by its range gives the highest recovery of the underlying cluster structure (Milligan & Cooper, 1988). Those findings are still used and supported within the academic field of cluster analysis (Gnanadesikan et al., 1995; Steinley & Brusco, 2008b).
Standardising by range:

$$\frac{\mathrm{x_{ij}}}{\max_j(\mathrm{x_{ij}}) - \min_j(\mathrm{x_{ij}})} \; for\ all\ \mathrm{x_{ij}} \tag{14}$$

Overall, the Minkowski distance function is favoured because it is easy and fast to compute, however before using it we have to standardise the attributes by dividing them by their range. If high linear correlation exists between variables, Minkowski distance can be distorted. In this case it is advisable to use Mahalanobis distance.

4.3.4. Clustering Techniques

In the academic literature there is no clustering technique that is universally applicable for all multidimensional data sets. Consequently, thousands of clustering techniques are presented in the academic literature with each built on different grouping criteria and similarity measures (Halkidi, 2001; Jain, 2010). Reviewing all clustering techniques would exceed our scope. However, articles like Xu and Wunsch (2005), Jain et al. (2000) or Berkhin (2006) provide a detailed overview and comparison of various clustering techniques.

Clustering techniques can be classified according to their basic approach of constructing clusters (see Figure 4.7). The main difference between clustering techniques is hierarchical and partitional clustering (Berkhin, 2006; Jain, Murty, et al., 2000; Xu & Wunsch, 2005). Within hierarchical clustering the algorithm can be mainly distinguished by its proximity measure which can be either single or complete-link based. Partitional clustering can be further divided into density, prototype, graph and mixture resolving based techniques. In the following we will briefly explain the hierarchical, the prototype as well as the mixture-resolving clustering technique by outlining their most common and fundamental clustering algorithms. Understanding the fundamental techniques automatically provides an introduction into many other clustering concepts as they mostly rely on the same assumption (Tan et al., 2005b). We leave out the graph-based clustering, which is often based on the minimum spanning tree, as it can be seen as a single-link hierarchical algorithm (Jain, 2010; Jain, Murty, et al., 2000). Furthermore, density-based clustering clusters based on dense areas within the data space (Berkhin, 2006; Tan et al., 2005b). As we will conduct cluster analysis based on a sample we cannot be sure that it resembles the density of the whole Netherlands. To minimise this risk, we focus on the other techniques. For a brief outline of graph-based and density-based clustering we refer to Appendix IX.

After the brief outline we summarise the characteristics of the algorithms and derive a framework for selecting suitable cluster algorithms based on that. Into this framework we incorporate recent

clustering techniques by analysing how they tackle drawbacks of the traditional algorithms. We limit ourselves to cluster algorithms which are available in official data analysing tools like SPSS or WEKA to ensure access and an easy application of our framework in practice.

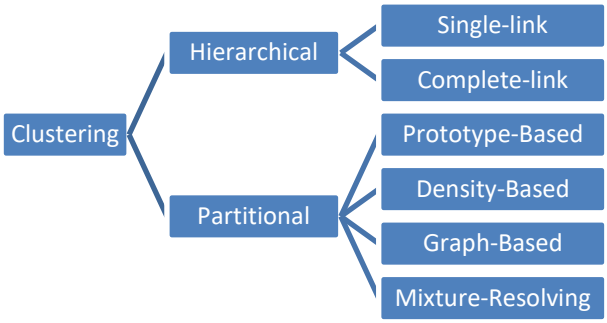


FIGURE 4.7: A CLASSIFICATION OF CLUSTERING APPROACHES

Hierarchical clustering approach

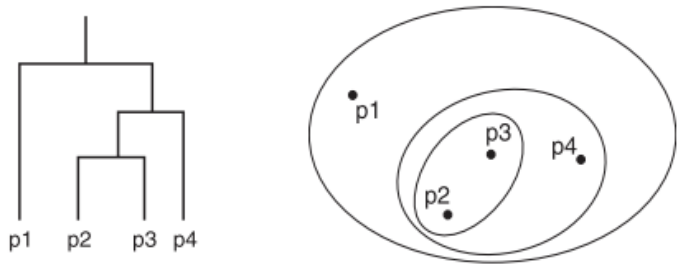
Hierarchical clustering provides a nested clustering in the form of a dendrogram (see Figure 4.8). In this research we focus on the most common and fastest approach to build this dendrogram, the agglomerative clustering (Tan et al., 2005b; Xu & Wunsch, 2005). It follows a bottom-up approach and starts with considering each single object as a cluster and iteratively combines clusters. It ends when all objects are within one cluster. In each step the closest pair of clusters is merged based on a certain notion of cluster proximity. With N objects in a cluster this step requires a computational time of $O(N^2)$ (Tan et al., 2005b).

The basic algorithm can be described as followed (Tan et al., 2005b, p. 516)

1: Computing the proximity matrix
2: **repeating**
  3: merging the closest two clusters
  4: updating the proximity matrix to reflect the proximity between the new cluster and the original clusters
5: **until** only one cluster remains.

Considering the algorithm we can see that merging decisions are final and local based on the proximity between clusters. There are two main principles to measure proximity: single-link, also referred to as nearest neighbour distance, where the proximity is the smallest distance between two objects of different clusters, and complete-link, also called farthest neighbour distance, where the maximum distance between two objects of different classes is used for measuring proximity (Xu & Wunsch, 2005). Both have their advantages and disadvantages. Complete-link results in compact and spherical clusters. Single-link clusters are often straggly, but more versatile and can detect chain like or concentric clusters. However, because of this versatility of cluster shapes single-link clustering is more affected by noise data. Thus, complete-link clustering should be applied to data sets with spherical cluster forms to avoid the risk of noise influence, whereas single-link is favourable for data sets that tend to have non-spherical clusters, but one should pre-process the data on noise and outliers (Jain, Duin, et al., 2000).

Finally, to gain disjoint clusters, we can make a horizontal cut through the final dendrogram at the desired dissimilarity level and consider each separated branch as a cluster. The same idea can already be incorporated in the algorithm by defining a stop criteria if a certain level of dissimilarity is reached (Jain, Murty, et al., 2000). However, Kettenring (2006), who examined applications of clustering algorithm in practice, found out that cutting the dendrogram with a horizontal line can be perilous as it might not capture the real cluster structure and that we still lack “sophisticated tools for extracting clusters from the dendrogram” (p. 21). Thus, defining accurate clusters in hierarchical clusters might be challenging.



(a) Dendrogram. (b) Nested cluster diagram.

FIGURE 4.8: HIERARCHICAL CLUSTERING OF FOUR POINTS (Tan et al., 2005b, p. 516)

Prototype-based

At prototype-based clustering each cluster is resembled by a prototype. Objects are assigned to the prototype with the highest similarity. Depending on the type of data, a prototype can have different forms. For continuous data, the prototype is often a centroid, which is the average of all objects within that cluster. If the data is categorical, the most representative point, the medoid, is selected as prototype. As the prototype is often placed in the centre of a cluster, prototype-based clustering is also referred to as centroid clustering. (Tan et al., 2005b) Another reference within the academic literature is "squared error-based"(Xu & Wunsch, 2005) as "the sum of squared error function is one of the most widely used criteria" (p.651) in that field. This criterion evaluates the clustering based on the distance of all objects to their assigned prototype, which is measured by the squared error.
The K-means algorithm is the most popular algorithm of all prototype-based clusterings as well as the most used clustering algorithm in scientific and industrial applications. K-means works with continuous data and defines the centroid as the mean of a cluster c_k. (Berkhin, 2006; Jain, 2010). The algorithm aims to minimise the sum of the squared error (SSE) over all K clusters (Jain, 2010):

$$SSE(C) = \sum_{k=1}^{K} \sum_{x_i \in C_k} \|x_i - c_k\|^2 \tag{15}$$

In contrast to aggregative hierarchical clustering solving this minimising problem is NP-hard, which means that it is computational infeasible to check all possibilities. Therefore, an iterative optimisation with a greedy algorithm is applied, which stops the algorithm if there is no significant change in the centroids (Berkhin, 2006; Jain, 2010).
The steps are as follows (Tan et al., 2005b, p. 497):

1: selecting K-points as initial centroids
2: **repeating**
  3: forming K-clusters by assigning each point to its closest centroid
  4: recomputing the centroid of each cluster
5: **until** centroids do not change.

Using this iterative process improves the positioning of centroids but does not guarantee obtaining a global optimisation as it converts to local minima. Therefore, k-means performs better if initial starting points are well chosen and clusters are well separated, meaning that the distance between clusters is larger than within a cluster (Steinley & Brusco, 2008a; Tan et al., 2005b). The academic literature suggests different methods for determining the initial centroids. One of the most common and easiest one is to run the k-means algorithm multiple times with different random initialisation and select the one with the lowest SSE (Jain, 2010; Steinley, 2006). Statistical software packages like SPSS offer default functions to determine the starting centroids, however Steinley (2003) shows that multiple random initialisations outperform those default functions.

Another drawback of this algorithm is that it requires a fixed number of clusters (K) (see Step 1), which is often not known in advance, however there are several approaches suggested by the

academic literature helping to determine a suitable number of clusters which we will present in this section later.

Step 3 requires a notion for measuring the distance. As the k-means algorithm requires to calculate the proximity between objects and centroids multiple times, proximity measures should be relatively simple (Tan et al., 2005b). The traditional k-means algorithm is based on the Euclidean distance, which can, especially in low dimensional space, quickly be calculated. In particular, k-means requires O(I*K*n*d) computation time, thus it depends on the number of required iterations (I), clusters (K), the objects (n) and attributes (d). However, using the SSE based on the Euclidean distance as criterion makes it difficult to detect suitable clusters, if the clusters have different sizes, densities or non-hyperspherical shapes (Tan et al., 2005b; Xu & Wunsch, 2005).
Overall, k-means is a simple and fast algorithm and efficient for data sets with hyperspherical and compact shapes. Furthermore, there are several approaches suggested by the academic literature helping to determine a suitable number of clusters and initial starting points, which we will discuss during the comparison later on.

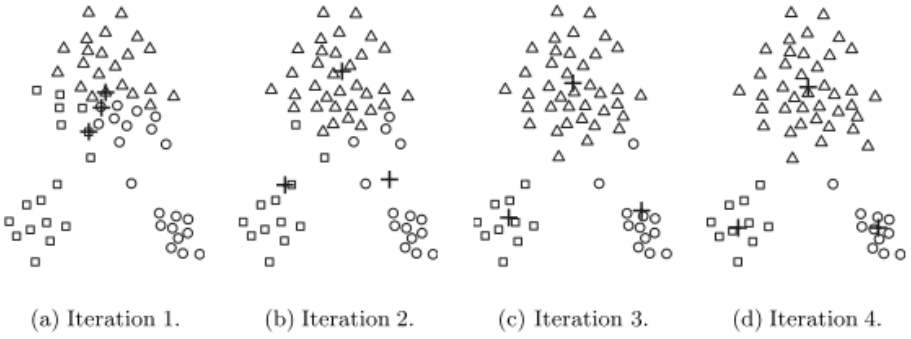


FIGURE 4.9: K-MEANS ALGORITHM WITH 3 INITIAL STARTING POINTS (Tan et al., 2005b, p. 498)

Mixture-resolving-based

Mixture-resolving-based, also called probability-distribution-based, clustering assumes that the data is independently drawn from different probability distributions. The probability distribution can differ by the type of density function or only by different parameters within the same function. Each distribution should at the end be resembled by a cluster (Berkhin, 2006; Jain, Duin, et al., 2000; Xu & Wunsch, 2005). A unimodal distribution would imply that the area around its mean forms a natural cluster (Berkhin, 2006). Thus, goal is to identify the different parameters of probability distributions. Therefore, the maximum likelihood estimation is used, which "considers the best estimate as the one that maximizes the probability of generating all the observations" (Xu & Wunsch, 2005, p. 653). For a detailed outline of the maximum likelihood estimator we refer to Larsen & Marx (2012). Mostly, the best estimator for a probability distribution cannot be identified analytically. Therefore, an approximation is required, and the most popular method to do so is the simple expectation-maximization (EM) presented by Dempster, Laird and Rubin (1977) (McLachlan & Krishnan, 1997 cited by Xu & Wunsch, 2005 and Berkhin, 2006; McLachlan & Peel (2000) cited by Steinley, 2006). The simple EM-model assumes that all objects follow a Gaussian density and are generated by a normal distribution. Thus, only the specific parameters, mean and covariance, for each Gaussian density have to be determined (Halkidi, 2001; Jain, Murty, et al., 2000). The algorithm follows an similar iterative process like the k-means algorithm (Celeux & Govaert, 1992): instead of improving the centroid position to minimize the SSE, the estimator is improved to maximise the likelihood.

The steps are as follows (Xu & Wunsch, 2005):

1: Selecting a parameter estimation
2: **repeating**
3: expectation step (E): computing the expectation of the complete data log-likelihood:
4: maximization step (M): selecting a new parameter estimate that maximizes the complete data log-likelihood.
5: **until** the convergence condition is satisfied

The output of this algorithm is an exact Gaussian probability distribution for each cluster, where each object is assigned with a certain probability to a cluster. Considering the algorithm we can see that, similar to the k-means algorithm, EM is highly sensitive to the initial parameter selection and might find only a local optimum instead of a global one. Furthermore, this algorithm does not rely on a distance measure and can thus be applied to heterogeneous data. The probabilistic foundation also implies the advantage that deciding on the number of clusters becomes a more tractable task. However, it has the drawback that it is sensitive to a singular covariance matrix (Xu & Wunsch, 2005). Finally, this algorithm not only provides cluster labels for each object, but also gives information on their probability distribution, which can be convenient when imputing other data points into the already existing clusters.

Comparison and Conclusion: The Selection Framework

In Table 4.4 we have summarised characteristics of the different algorithms. For the sake of completeness we have also included the density-based cluster algorithms, DBSCAN and OPTICs, as they offer many advantages which the other lack. We will compare 5 aspects: input, cluster shapes, outlier, time complexity and output. Firstly, the required *input* for cluster algorithms is not only often used to classify them (Jain, 2010), but also critical when assessing the suitability for a particular problem context as in practice often not all information is obtainable. Secondly, we will compare the favourable *cluster shapes* of the algorithm as the data shape has an enormous impact on the outcome of the algorithm (Halkidi, 2001; Jain, Murty, et al., 2000). Thirdly, we will evaluate the degree to which they can handle *outliers*. Fourthly, we compare the *computational complexity*, which is "a measure of how many steps the algorithm will require in the worst case for an […] input of a given size" (L. Hall, 1996, p. 95). The computational complexity will not only indicate the required time, but also if it can handle larger data sets. Finally, we will assess the *output type* of the algorithm, which should fit with the aim of the clustering.

Whereas the first three aspects mainly emphasise that the output of an algorithm highly depends on the data characteristics and input parameters values (Halkidi, 2001), the remaining two are more relevant from a practical perspective.

As there is no universal applicable cluster algorithm, we have created a flowchart based on the characteristics specified in Table 4.4 to provide a guideline for selecting based on the specific situation a suitable algorithm (see Figure 4.10). The final selection of cluster algorithm also depends on the application need, for instance if a partitioning of the whole data set is required or only to identify the most cohesive clusters (Jain, 2010). One of the main challenges is to determine the number of clusters. Therefore next to this framework we finalise this section by presenting briefly different approaches of the academic literature for tackling this problem. As we want to provide a practical guideline, we will keep the focus on those algorithms that are accessible in data mining tools like WEKA.

Framework for selecting a suitable clustering technique

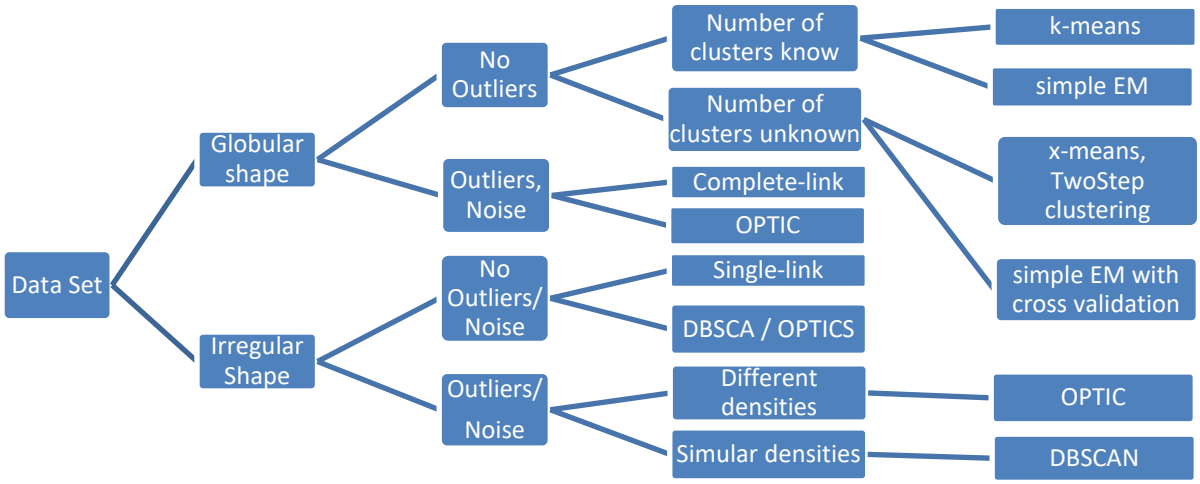


FIGURE 4.10. FRAMEWORK FOR SELECTING A CLUSTERING ALGORITHM

The quality of the output of a clustering algorithm highly depends on the assumption of the data shape that the algorithm incorporates. Cluster algorithms find and create almost always clusters even if the intrinsic clusters of the data set does not fit the assumptions of the algorithm (Tan et al., 2005b; Xu & Wunsch, 2005). Therefore, it is critical to select only those cluster algorithms that coincide with the cluster tendency of the data.

If the cluster shapes are irregular, k-means and complete-link aHC are unsuitable as they both tend to form hyperspherical shapes and should therefore only be applied in case of data sets with globular shapes (Jain, Murty, et al., 2000; Kettenring, 2006; Tan et al., 2005b; Xu & Wunsch, 2005). The traditional k-means method is based on Euclidean distance measure which tends to create spherical or ball-shaped clusters. K-means it not limited to Euclidean distance, however, the proximity measure for k-means should be kept simple, because it calculates the proximity between objects and centroids multiple times. Therefore, more complex proximity measures would come with high computational expenses (Jain, 2010; Tan et al., 2005b). A detailed description of the proximity measures and their implications for cluster forming is given in Section 4.3.3.

Complete-link measures proximity based on the maximum distance between any two objects of two different clusters. Thus, consequently all the other objects of the clusters are at least as close to the new connected cluster as the object with the maximum distance. This results in highly compact clusters but also tends to form hyperspherical shaped clusters (Tan et al., 2005b). Another advantage of complete clustering is that following this approach is less affected by outliers. In contrast to that, k-means clustering considers every object, being irrelevant if they are noise or outliers, for determining the optimal centroid position and is thus highly affected by them. Therefore, even though k-means clustering is with O (INkd) much faster than the complete-link clustering with O (N^2), one should use complete-link clustering if many outliers and noise exist.

If outliers and noise data can be mostly eliminated during the data prepressing and clusters tend to have a hyperspherical shape, k-means can work very fast and efficient (Jain, 2010). Depending on the required output, one can also select a simple EM-clustering. Simple EM provides the probability distribution per cluster rather than a prototype. However, to apply the k-means or the EM algorithm, we have to select the right number of clusters and suitable initial centroids. In the last part of this section we will analyse the solutions proposed within the academic literature.

If we assume an irregular cluster shape, we should rather select a density based approach. The DSBCAN algorithm can detect arbitrary shapes and is quite robust against outliers. The drawbacks are that it is highly sensitive to input parameters (radius and minimum number of objects) and has difficulties to handle different densities between clusters (Tan et al., 2005b). A less sensitive algorithm, which can cope with different densities, is the Ordering Points To identify the Clustering

Structure algorithm (OPTICS) presented by Ankerst et al. (1999) (Jain & Dubes, 1988). However, instead of a clustering, OPTICS plots the distance of each object in a reachability graph (see Figure_Apx IX-4). By that it visualise the density allocation of objects within the data space. Thus, overall OPTICS can be good tool for assessing the cluster structure, however it does not provide a final clustering. For an exact outline of the algorithm see Ankerst et al. (1999). For all density based clustering accounts that the application to only a sample of data is risky, because the densities might not be representative.

Another algorithm which can be used for arbitrary shapes is single-link hierarchical clustering. In contrast to DBSCAN it requires no input, however has with O (n^2 log n) compared to O (n log n) a higher computational time. Another drawback is that it is quite sensible to outliers and noise, thus pre-processing the data is critical.(Jain, Murty, et al., 2000) One approach to identify and eliminate outliers is to make an initial run with the hierarchical cluster algorithm and use the resulting agglomerative table. This table shows the sequence of merging objects and the distance between objects and clusters at the moment of merging. Given the hierarchical cluster algorithm, which merges the closest object first, objects that are merged at the end tend to have a high distance to other objects and therefore should be seen as potential outlier (Ketchen & Shook, 1996).

Overall, the presented flowchart is not a definite rule to follow, but still requires critical assessment at each step of the clustering analysis. Each clustering technique has its own advantages and disadvantages. Therefore, we advise to make use of the different advantages by combing different clustering techniques. For instance, OPTICS can be used preceding the k-means algorithm. By first visualising the cluster tendency in the reachability-plot, we can estimate the number of clusters and use it as an input for k-means. Furthermore, the quality of the cluster outcome can be improved by pre-running our data set with the hierarchical clustering technique, where we can identify and eliminate outliers. Subsequently, the processed data set can be used as an input for the same algorithm or be applied to others.

In case the decision on cluster shape or outliers is not clear, we advise to use multiple clustering techniques as well. By that we can ensure that all possibilities are covered. For instance, in case of uncertainty on the cluster shape, we could apply k-means (covering globular shape) as well as single-link hierarchical clustering (covering straggly shape). Subsequently, the cluster outcome which performs best based on the on validation, should be selected.
Before defining the validation criteria in Section 4.3.5, we will briefly outline techniques to determine the number of clusters.

Approaches to apply cluster algorithms without k as input
Within academic literature there are many solutions for determining the number of clusters. For instance we can assess the cluster tendency by visualising the data structure, which however is limited to 1 till 3 dimensional objects (Halkidi, Batistakis, & Vazirgiannis, 2002a). The most common method is to conduct a sensitivity analysis by running k-means multiple times with different K and to select the best one based on a predefined criterion (Jain, 2010). Two common criteria are the Bayes Information Criterion (BIC) and the Akaike Information Criterion (AIC) (Burnham & Anderson, 2004; Xu & Wunsch, 2005). Given a set of free parameters, which is in our case the number of clusters, both criteria assess the increase in the likelihood function when increasing the parameter. However, to prevent an overfitting of the parameters, thus to prevent to add parameters with a poor improvement in cluster quality, a penalty term for the number of parameters is used. Both, AIC and BIC, have their own penalty system and theoretical targets (see Appendix VII), and both identify adequate models for the data set given their own theoretical targets. Thus we cannot say that one performs always better than the other and hence there is no overall rule on their application (Kuha, 2004). Kuha (2004) suggests to use both criteria to ensure the robustness of the choice.

There are two algorithms automatically defining K by applying BIC and/or AIC, which are x-means (WEKA) and the TwoStep clustering (SPSS). Pelleg and Moore (2000) present the so called x-means

algorithm which incorporates the BIC as splitting criteria. X-means follows two main steps iteratively until the maximum number of k is reached; subsequently they select the configuration with the best global BIC score. The first step, the improve-params, is to run the conventional k-means to convergence. In contrast to k-means, they incorporated an improvement step as second step, which splits each centroid (=parent centroid) into two children centroids and reruns the k-means locally given all objects assigned to the parent centroid. Depending on the model selection criterion, BIC which is measured locally as well, either discards the parent or children centroids (for a detailed explanation see Pelleg & Moore (2000)). Their study with 4 dimensional data showed that it performs better and faster than manually running and testing k-means, but also suggests further research if AIC can be incorporated as well.

The statistical software tool, SPSS, incorporates the BIC and AIC criterion in their TwoStep clustering algorithm as well to determine in combination with the hierarchical clustering algorithm the number of clusters. Therefore, for each possible number of cluster BIC or AIC is calculated to give a first estimate of the exact number, subsequently hierarchical clustering with a centroid linkage (distance between clusters is represented by the distance between their centroids) is applied to determine the exact clusters based on the greatest change in distance between merged clusters (SPSS, 2001).

Another approach used to determine the number of clusters without BIC or AIC is to conduct a cross-validation with simple EM offered by WEKA. Cross-validation takes part of the sample as trainings set and subsequently tries to predict the rest of the data set. Multiple iterations are conducted, with a new randomly selected trainings sample in each iteration. Over all those iterations the average likely-hood is calculated; starting with k=1 the first average loglikelihood is calculated and as long as it improves by a certain minimum threshold, set by the user, the number of cluster increases as well. In contrast to AIC and BIC, there is no penalty criterion for increasing the number of clusters. Similar to AIC and BIC it is a model-based criterion and all of them, incorporated in WEKA and SPSS, assume that the attributes are independent and normally distributed (Halkidi, 2001; Jain, Murty, et al., 2000). However, according to IBM (2016) the TwoStep clustering is by empirical internal testing shown to be fairly robust to violations of those assumptions. Nevertheless, one should always use common sense and model selection techniques only as a guideline (Agresti, 2013).

Instead of using automatic algorithm which relies on a normal distribution of the cluster attributes, we can also perform it manually by running the clustering multiple times with different k numbers of clusters. For each outcome predetermined internal criteria are measured. By plotting the results in a graph, we can observe the change by increasing the number of clusters. The aim is to determine the maximum (minimum) of the plot in order to identify the optimal number of clusters. However, for criteria that increase with the number of clusters, we search for the significant local change, the so called “knee” in the plot (Halkidi, Batistakis, & Vazirgiannis, 2002b; Jain, Murty, et al., 2000; Ketchen & Shook, 1996; Tan et al., 2005b). If there is no knee, it may indicate that there are no natural clusters within the dataset (Halkidi, 2001). The criteria used are the same as for internal validation, which we will outline in Section 4.3.5, as those indicate similarity and compactness of clusters. Most commonly the mean SSE is used for k-means for measuring the compactness within clusters, combined with the mean silhouette coefficient to indicate the cluster separation (Jain, Murty, et al., 2000; Tan et al., 2005b). For hierarchical clustering, the “knee” approach can be used by plotting the number of clusters against the agglomeration coefficient as a significant change in the coefficient indicates that very dissimilar clusters are combined (Ketchen & Shook, 1996).

TABLE 4.4: MAIN PROPERTIES OF THE CLUSTERING ALGORITHMS

Cluster algorithm	Input	Data Type	High dimensional Data	Complexity (I = number of iterations, n = number of objects, k = number of clusters, d = number of dimensions/attributes)	Cluster Shape	Resistant to Outliers	Output
Ag. Hi. Single-link		Numerical	No	$O(n^2 \log n)$ (time)	Straggly shape	No	Dendrogram
Ag. Hi. Complete-link		Numerical			Hyperspherical shape	Medium	Dendrogram
k-means	- Number of clusters	Numerical	No	O(inkd) (time)	Hyperspherical shape	No	Centre of clusters
DBSCAN	- Radius (Esp.) - Minimum number of objects (MinPts)	Numerical	No	O(n log n) (time)	Arbitrary shapes	Yes	Assign objects to cluster
OPTIC	- Radius (Esp.) - Minimum number of objects (MinPts)	Numerical	No	O(n log n) (time)	Arbitrary shapes	Yes	Reachability-plot
Simple EM	- Number of probability distributions	Numerical	No	O (indk) (time)	Hyperspherical shapes	No	Probability distribution per cluster

4.3.5. Cluster Validation

Cluster analysis is an unsupervised process, which contains no universal solution approach. The quality of an algorithm is highly sensitive to the attributes of the data set (e.g. geometry, density) and the input parameters (Halkidi et al., 2002a; Jain, 2010). As there are no golden standards, both factors are usually subjective derived and often require a trial-and-error procedure (Jain, Murty, et al., 2000). To determine the meaningfulness of the (trial-error) output, validation is critical as it is an objective assessment (Dubes, 1993). It is also possible to let the output be assessed by an expert. However, this requires a clustering being easy to interpret and containing a good visualisation. Thus, this is often limited to low dimensional data but also depends on the type of clustering technique applied. For instance mixture-resolving model provides a clear description of the probability distribution per cluster, k-means provides a prototype per cluster or the hierarchical clustering gives a dendrogram where we can see how the attribute values are spitted, thus all can easily be understood. However, the output of density based clustering provides only cluster labels and is therefore harder to interpret (Berkhin, 2006).

Considering the more objective and quantitative validation, we can distinguish three main criteria: internal, external and relative criteria (Berkhin, 2006; Halkidi et al., 2002a; Jain, Duin, et al., 2000; Tan et al., 2005b). In the following we will discuss each of those criteria and provide different methods for assessing them.

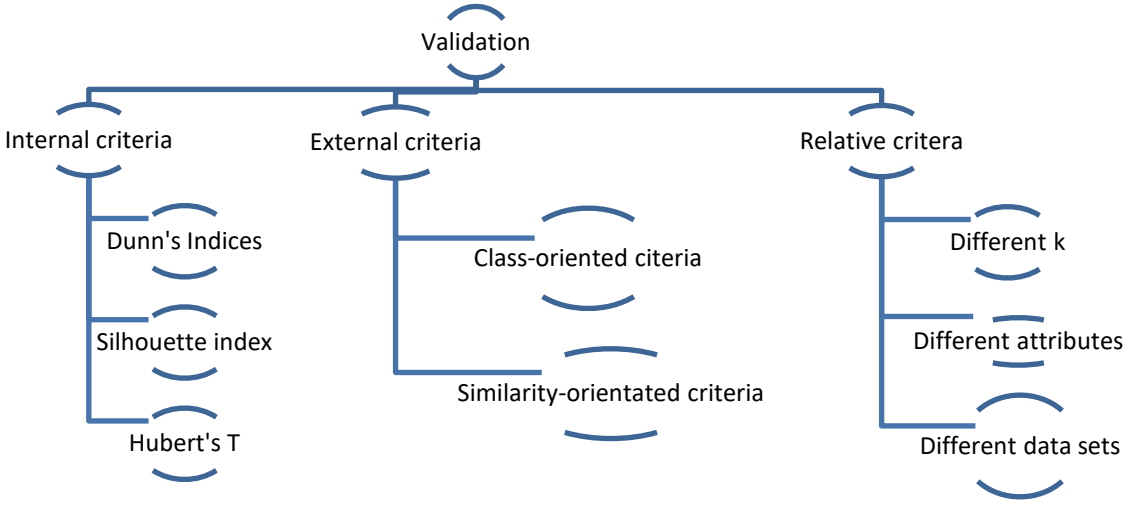


FIGURE 4.11: VALIDATION CRITERIA

Internal criteria

Internal criteria solely rely on information of the data set to evaluate the cluster outcome. There are three aspects we can measure based on internal information: the cluster compactness and separation, the fit of the object and its cluster and finally the correlation between the ideal similarity matrix and the distance matrix.

Cohesion and Cluster Separation

The goal of clustering is to generate high similarity within a cluster and high distinction between clusters, which is often evaluated based on cluster cohesion and separation. Cluster cohesion (CC) depends on the compactness and tightness of each cluster (C_i, with i=1, …k) and indicates how similar objects within a cluster are. Therefore, we can measure the average distance between the objects (x,y) within one cluster (CC_1). In case of a prototype based clustering, we can also use the total proximity of all objects to its cluster centroid (c_i) as cohesion measure (CC_2) . This is identical with the SSE when the proximity is measured by the squared Euclidean distance (see Formula 15). (Tan et al., 2005b) In order to evaluate the overall cohesion of a clustering, we can take the sum of the coherence measures over all K clusters.

$$\mathrm{CC1}_1(\mathrm{C}_i) = (\mathrm{l/M}) \sum_{x,y \epsilon C_i} proximity\,(x,y) \tag{16}$$

$$\mathrm{CC}_2(\mathrm{C}_i) = \sum_{x,y \epsilon C_i} proximty\,(x,c_i) \tag{17}$$

Cluster separation in contrast indicates the distinctiveness and separation of one cluster with the remaining clusters. Therefore, we have to define the linkage methods used for measuring the distance (d) between two clusters. For prototype-based clustering Tan and Steinbach (2005b) suggest the centroid linkage, which represents the distance between clusters by the distance between their centroids. Dunn (1974) in contrast proposed to use the single link distance, by taking the smallest distance between any object of C_i and C_j as distance measure (Halkidi, 2001). Overall, various linkage methods can be applied, which depends on the preference of the user (complete, average, average to centroid etc.) (Brun et al., 2007).

An often used indicator for measuring the compactness and separation of clusters is the **Dunn index**, proposed by Dunn (1974) (Brun et al., 2007; Halkidi, 2001), which gives the ratio between the maximum distance between two clusters and the cluster diameter:

$$D = \min_{i=1,\dots K} \left(\frac{d(C_i\,, C_j)}{\max_{k=1\dots K} diam(C_k)} \right) \tag{18}$$

with

$$d = \max_{x\,\epsilon\,C_i, y\,\epsilon\,C_j} d(x,y)$$

$$diam(C) = \max_{x,y\,\epsilon\,C} d(x,y)$$

The idea is that if clusters are well separated, the distance between them (d) is high, and if a cluster is cohesive its diameter is relatively small. Therefore, the higher the index, the more distinct and cohesive the clusters are. The drawback of this index is the high vulnerability to outliers as a single one could result in an increase of the diameter. Furthermore, the computational time is quite high. A more simplified indicator is the similarity measure based on the ratio between cohesiveness of clusters and separation between them (Halkidi, 2001; Steinley, 2006; Tan et al., 2005b), which is known as the **Davies-Bouldin (DB) index**: For two clusters the similarity can be calculated as follows:

$$SI_{i,j} = \frac{\mathrm{CC}_i + \mathrm{CC}_j}{d_{i,j}} \tag{19}$$

with $d_{i,j} = dissimilarity\ between\ two\ clusters\ which\ can\ be\ represented\ by\ d(C_i\,, C_j)$

The DB index, which is an overall indicator for all clusters, is then defined as:

$$DB = \frac{1}{k} \sum_{k} SI_i \tag{20}$$

$$with\ SI_i = \max_{i=1,\dots k, i \neq j} SI_{ij}$$

Overall, we can see that the DB index is an average of the similarity between all clusters. The highest similarity possible is one, however as the goal of clustering is to create distinctive clusters, the aim is to create clusters with a DB close to zero.

Instead of analysing the cohesion and the separation on a cluster level, we can also evaluate it for each object, which we discuss in the following.

Fit of an object and its cluster

The goal of clustering is to group similar objects together. In order to assess this, we can evaluate the fit of the object with its current cluster compared to the fit with other clusters. Therefore, the **Silhouette Coefficient** is a common applied measure introduced by Rousseeuw (1987) (Berkhin, 2006; Tan et al., 2005b) and can be calculated as follows:

1. calculate a_i ,the average distance of object i to all objects in its cluster

2. calculate the average distance of object i to all objects for each cluster that does not contain i. Determine b_i by taking the minimum of all averages.
3. calculate the Silhouette Coefficient:

$$SC_i = (b_i - a_i)/\max(a_i, b_i) \tag{21}$$

The Silhouette Coefficient has a range from -1 to 1. The aim is to have a low a_i indicating high similarity between the object and the other objects of its cluster. A high b_i indicates that the object does not fit in other clusters. Thus overall, the closer SC_i is to 1 the better. In contrast, a negative SC_i $(b_i < a_i)$ indicates that the object would fit better in another cluster than the current one, thus we should reconsider the cluster assignment. Finally, it is also possible to evaluate the overall clustering by taking the average of all SC_i's, which indicates the overall cohesion and separation of the clustering (Tan et al., 2005b). An average Silhouette Coefficient higher than 0.5 indicates a good clustering, from 0.2 to 0.5 a fair clustering and lower than 0.2 a poor clustering (IBM, 2016).

Overall, both the average Silhouette Coefficient and the Davies-Bouldin (DB) index indicate the compactness and separation of the clustering. The Silhouette Coefficient can automatically be calculated by the statistical software SPSS, requiring only the cluster membership of each object as input. In contrast, Davies-Bouldin (DB) requires to define and calculate the cohesion measure (CC) as well as the distance measure between clusters. Thus, for a practical approach Silhouette Coefficient is easier and faster to apply and therefore more advisable in our case.

Correlation between the ideal similarity matrix and the distance matrix

Ideally if objects are within the same cluster, they should show a similarity of 1, while they should have zero similarity with objects of other clusters. Based on this we can set up an ideal similarity matrix D with each row and column representing one object and values of 1 and 0 indicating if they are in the same cluster or not. Besides to that we create the actual similarity matrix P by calculating the similarity between each pair of objects. Subsequently, we can calculate the correlation between those two matrices Γ, also known as Hubert's correlation. (Halkidi et al., 2002a; Tan et al., 2005b)

$$\Gamma = \frac{1}{M}\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} D(i,j)P(i,j) \tag{22}$$

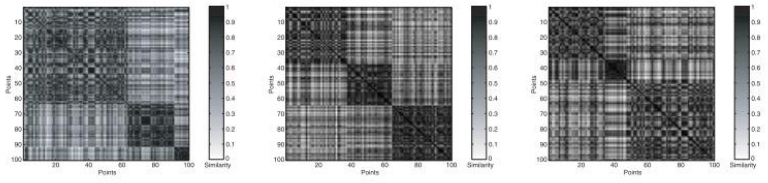
$$with\ M = \frac{n(n-1)}{2}\ (number\ of\ pairs\ of\ different\ objects)$$

The higher the correlation, the closer the cluster output is towards the ideal. Hubert's correlation can also be applied with small adaption to a dissimilarity matrix (which is often the distance measure between objects) instead of the similarity matrix (Tan et al., 2005). As seen in 4.3.4 often cluster algorithms are based on the Euclidean distance measure. Frey and Dueck (2007) propose to use the negative Squared Euclidean distance as similarity measure, where the similarity between object i and object j can be defined as:

$$s(i,j) = -\|x_i - x_j\|_2^2 \tag{23}$$

$$with\ \|x_i - x_j\|_2^2 = squared\ Euclidean\ distance$$

We can also test the correlation visually by ordering the actual similarity matrix based on the clusters and indicating the similarity based on a colour scale (from light to dark). Subsequently we plot it with the ideal outcome of a block-diagonal matrix where each block and block size represents a cluster and its size. If the similarity is high we expect one overall dark block, however the more the colour within one block varies, the weaker the cluster is. (see Figure 4.12)(Tan et al., 2005b)



(a) Similarity matrix sorted by DBSCAN cluster labels.

(b) Similarity matrix sorted by K-means cluster labels.

(c) Similarity matrix sorted by complete link cluster labels.

FIGURE 4.12: VISUALISATION OF THE SIMILARITY (2005b, p. 543)

Overall, comparing the ideal with the actual outcome helps to evaluate the quality of the outcome with the main advantage that it can be applied to all cluster algorithms. However, it comes with high computation time as we have to calculate the distances or similarities for all pairs (O (m^2)). SPSS provides a proximity matrix incorporating the pairwise sqaured euclidean distance between all objects (IBM, 2016) and by using the similarity measurement proposed by Frey and Dueck (2007) it is easy to gain the required information.

Finally, the visual approach is useful if we want to get more insight into one cluster outcome, however it does not enable a good comparability between different kinds of clusterings. In contrast, the Hubert's correlation coefficient allows a quantitative comparison between clusters, but has the drawback of high computational complexity given the Hubert's correlation formula. Therefore, it is not advisable for data sets containing many objects.

External Criteria

External criteria evaluate the cluster analysis based on external information. There are two main approaches: the class-orientated approach which determines class labels based on externally known results and compares the fit between the predicted class labels and the actual class label. For this evaluation common classification-orientated measures can be used, such as purity, entropy and the F-measure (Berkhin, 2006; Tan et al., 2005b). Classification-oriented measures require much information; not only on the kind of clusters that we expect, but also an estimation on which object belongs to which cluster. This information is often not obtainable when conducting a cluster analysis as we otherwise would rather conduct a classification (supervised learning).(Tan et al., 2005b) Therefore, we do not focus on those validation techniques, but for more information we refer to Tan et al. (2005a).

The second approach is similarity-orientated. It only requires an assessment, if a certain pair of objects is likely to belong to the same class. Based on the clustering outcome we can also give the information, if a certain pair of objects belongs to the same cluster or not. The most common measures used to indicate the fit between this external information and the cluster outcome are the Rand statistic measure, the Jaccard Index and the Fowlkes- Mallow Index. Comparing both partitions, the externally based and the one based on the clustering, we can define following variables:

a = number of pairs where the objects belong to the same class and the same cluster
b = number of pairs where the objects belong to the same class, but to different clusters
c = number of pairs where the objects belong to different classes, but to the same cluster
d = number of pairs where the objects belong to different classes and to different clusters
M= a+b+c+d = max. number of pairs within a data set with N objects= $\frac{N(N-1)}{2}$

The Rand Statistic R gives the proportion of matches between the external information and the cluster outcome compared to the mismatches:

$$R = \frac{(a+d)}{M} \tag{24}$$

The Jaccard Index measures the proportion of pairs belonging to the cluster and class compared to all pairs belonging to the same group in at least one of the partitions:

$$J = \frac{a}{a+b+c} \tag{25}$$

Finally, the Fowlkes-Mallow Index compares the geometric means of proportion of pairs that belong to the same group in both partitions vs. pairs that belong to the same group in each partition.

$$FM = \sqrt{\frac{a}{a+b} * \frac{a}{a+c}} \tag{26}$$

Each of the three measures focuses on different kinds of similarities and thus should all be used to evaluate the cluster outcome. For all three holds: the higher the indices, the greater the fit between cluster outcome and external information. The range of the Rand Statistic and the Jaccard Index is between 0 and 1, thus the closer the measures are to 1 the better the quality of the clusters. The drawback of those tools is that it requires high computational complexity depending on the number of objects (N) which makes it sometimes not feasible, for instance the Rand Statistic has a complexity of O (N^2). To minimise the complexity, one can base the test a sample of objects. (Berkhin, 2006; Halkidi et al., 2002a; Tan et al., 2005b)

Both approaches are quantitative, however, for a validation one can also use an expert from the object domain to gain a qualitative assessment of the clustering. Therefore, it is important that the clusters are easy to interpret and in the best case can be visualised (Becher, Berkhin, & Freeman, 2000).

Relative Criteria

Relative criteria evaluate the consistency of a cluster algorithm by comparing the obtained clusters determined by the same algorithm under different parameters. The most common parameter tested is the number of cluster and the cluster attributes, but also the data set. Similar to a sensitivity analysis it is important to change only one of the parameters in order to evaluate the impact of the change. The impact of the change on the quality of the clustering can be measured with the internal and external validation criteria. (Brun et al., 2007; Halkidi, 2001; Tan et al., 2005b)

We have already discussed specific approaches on how to compare the cluster outcome in order to select a suitable set of attributes in Section 4.3.2. In Section 4.3.4, we have discussed different algorithms for finding the right number of clusters.

Overall, all validity measures are indicators for the quality of the clustering output. However, there are no clear rules that tell whether a clustering is significant good or bad. Some validity measures have certain minima and maxima, which we can use as an indicator for the significance (Tan et al., 2005b). Often however, no fixed ranges or thresholds are given, which makes it even harder to interpret. In that case relative comparison can, to some degree, help to evaluate the validity outcome. As clustering is data-driven and is used for exploratory means, evaluating the statistical significance can often not be used, but for evaluating the relatively difference of individual variables within or between clusters using the ANOVA can serve as a good practical tool (Kettenring, 2006).

4.4. Conclusion of the Literature Review

At the beginning of this chapter, we have discussed and defined the benchmarking process. Benchmarking is a continuous process of collecting, analysing and assessing performance measures with the benchmarking partners in order to find best practices or to identify gaps for improvement. To actually gain a process improvement, we have to ensure to not only conduct discussions, but also to determine action plans for implementing and monitoring the learnings. Therefore, we have defined the steps for the benchmarking that PostNL should follow in Table 4.1.

Subsequently, we have determined a framework for deriving suitable performance measures. Different frameworks exist, however the four perspectives on operations strategy (Slack et al., 2010) fits best to derive performance measures for the mail delivery process. To ensure high quality performance measures we have defined recommendations aligned to the requirements of the benchmarking model of PostNL. Moreover, we presented the performance measure record sheet of

Neely et al. (1997) as a practical tool to ensure that performance measures fulfil those recommendations (see Table I.2). It is essential not only to consider the purpose of the performance measure, but also, if the data is available on the right level, to calculate them.

Finally, we have defined the cluster analysis process. The literature on cluster analysis is quite complex and derived various approaches for the different steps within cluster analysis: selecting and weighting attributes, selecting a clustering technique and validating the outcome. Within the current literature there are articles like Halkidi (2001) or Tan et al. (2005b) summarising cluster analysis. However, many are not useful as a guideline for practical applications of cluster analysis; either the presented methods do not cover all steps of the clustering process (see Figure 4.3) (for instance Halkidi (2001) with no information about attribute selection and weighting) or do not provide a clear framework on when to use them (for instance Tan et al. (2005)).

Defining clear guidelines for clustering is challenging, because clustering is data-driven and is used for exploratory means and therefore there is no clear right or wrong. However, conducting the literature review, we have recognised that each method or technique requires different information and patterns within the data. Making use of this, we have been able to define two frameworks for selecting weighting approaches and clustering techniques (see Figure 4.6 and Figure 4.13). For each framework we provided an explanation that gives an understanding on why the models have certain advantage and disadvantages, but it is still easy to follow for people who are not technical experts. Those frameworks make cluster analysis more accessible and provide an easier application to practical problems. Providing easy access and applicability is critical given the raise of big data (Halkidi, 2001) and the need of companies like PostNL to explore it.

However, it has to be understood that clustering is not always useful - only if there exist intrinsic groups among the cluster objects, because even if we select important attributes for defining a cluster object, it does not mean that they support the same cluster tendency. Therefore, each step of the clustering is essential; before applying a clustering technique, we have to analyse and to assess the cluster attributes and the cluster tendency. Afterwards, we have to validate it by internal, external as well as relative criteria. The drawback, however, is that within the field of clustering those criteria have often no clear thresholds that they have to reach in order to be satisfied.

In the following chapter, we will apply the findings of our literature research to the benchmarking model of mail delivery process.

5. Developing the Benchmarking Model for the Mail Delivery Service of PostNL

In this chapter we will combine the findings of previous chapters to develop the benchmarking model for mail delivery service. Therefore, we will follow the steps defined in Section 4.1. However, considering our research scope not all steps are covered. Steps 1 up to 7, which concern the development of the benchmarking model, are fully covered in this research. As our research scope (see Section 1.5) focuses on the clustering, we only cover Step 8 on data collection partly by analysing the data availability and the implication for our selected performance measure and clustering. Step 9 up to and including Step 13, which concern comparing, defining, implementing and monitoring best practices, are not covered in this research as we do not implement and perform a full benchmarking.

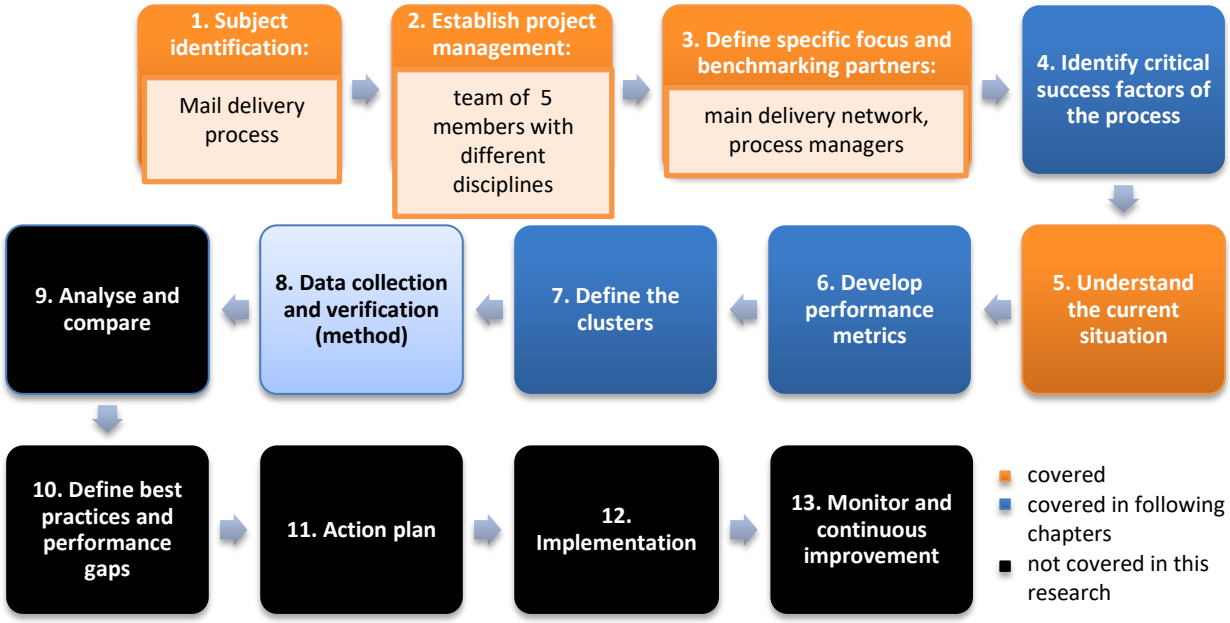


FIGURE 5.1: BENCHMARKING PROCESS FOR THE MAIL DELIVERY PROCESS OF POSTNL

Looking at the benchmarking process, we have already finished the first three steps and Step 5. The first step, the benchmarking subject, is defined as the national mail delivery process in our research scope (Section 1.5). The second step, the project management, incorporates a project team as well as a project planning. Our project team consists of critical stakeholders for the benchmarking model, which includes the senior controller, the senior process manager of optimisation, a process manager of delivery and, for the expertise knowledge, a senior manager of logistic strategy. The project planning with specification of the milestones is developed in consultation with the project team. The third step, the benchmarking focus and benchmarking partners, is defined in the research scope (Section 1.5). We focus on the main mail delivery network with the starting point at the HUBs and depots and the end at the delivery of the mail at the customer, retailer or public mailbox of PostNL. Benchmarking partners are the process managers of the delivery areas. Finally, Step 5, understanding the current situation, is covered in Chapter 2**,** where we have analysed the mail delivery process in detail by outlining each step of the process and clarifying the factors that might influence the performance.

The next step, that we have to conduct, is the identification of the critical success factors, which we do in the following section. Based on this we can define performance measures in Section 5.2, which helps us to derive potential attributes in Section 5.3. Finally, in Section 5.4 we evaluate the information requirement and the availability for the cluster attributes.

5.1. Defining the Critical Success Factors

Regarding the literature review we conclude that the four perspective framework of Slack et al. (2010) is a suitable guideline to derive critical success factors (CSFs) for the mail delivery process. Therefore, we analyse the mail delivery process from four perspectives: top-down, bottom-up, market requirement and operations resources perspective.

The top-down perspective considers the expectations of the business on the operations. Therefore, we analyse the strategy for mail delivery defined by the logistic strategy department.
The bottom-up perspective considers the day-to-day operations to determine CSFs. Therefore, we compare theory and practice, by determining the key elements' suggested by the academic literature first (Chan et al., 2009; Landeghem & Persoons, 2001; Slack et al., 2010) and subsequently comparing it with the findings based on the interview with the line employees (postmen, team leaders, process managers) (see Section 3.1).
The market requirements perspective considers "what the market position requires operation to do" (Slack et al., 2010, p. 65). It depends on customer needs and competitors' performance which we assess with the importance-performance matrix (Slack et al., 2010).
Finally, the operational resource perspective considers the capabilities of operational resources and is based on the resource-based view (RBV). RBV says that a firm could create a sustainable competitive advantage based on the core competences of its resources (Johnson, Whittington, & Scholes, 2011). Thus, to assess the performance of mail delivery process we clarify (see Chapter 2) PostNL's main resources with the capabilities but also their constraints based on the observations and the process analysis.

By analysing those four perspectives we are able to create a full picture of the areas that are critical for the success. In Table 5.1 we have summarised the CSFs from each perspective and listed it based on its importance. For a detailed analysis we refer to Appendix VIII.

TABLE 5.1: SUMMARY OF CSF OF THE MAIL DELIVERY PROCESS

Top-down	Bottom-up	Market requirement	Operational resources
Flexibility	Customer satisfaction	Cost	Postmen efficiency
Cost efficiency	Quality	Delivery time	Efficient usage of depots
	Flexibility	Quality	Efficient usage of means of transportation
	Employee commitment		
	Cost		

We can see that each perspective has its own priority; however, there is a high overlap between the named CSFs:

Costs
Costs mean producing cheaply in order to offer a reasonable price for the market and still gaining reasonable profit. The mail delivery process is one of the major cost components of the total costs. Considering the current market position of PostNL towards its main competitor Sandd, Sandd can offer a lower price than PostNL on average. Looking at the market, customers select rather on price than on delivery time (see Appendix VIII). Consequently, PostNL is losing some customers to Sandd, which can be seen based on the increasing market share of Sandd. While it had a market share of 25-30% in 2014, it increased to 30-35% in 2015 (ACM, 2016).
However, one should not only assess costs isolated, but also in relation to the output. Chan et al. (2009) do not only use costs for the performance in the postal industry (manufacturing, item carrying, storage and distribution), but also the return on assets as a measure of efficiency in utilizing assets. The objective is to enhance productivity at a competitive price and quality. PostNL highly focuses on costs in its MJ dashboard; however it does not consider the utilization of assets even though part of their strategy is to enable a cost efficient delivery process by an optimal usage of different means of transportation. From the operational perspective we can confirm that assets

utilization in the form of optimal use of depots and means of transportation is critical in order to minimize the costs.
Therefore, we can conclude that costs are a critical element of mail delivery process, however not having been implemented by PostNL in the right way until now.

Flexibility
Flexibility means the ability to change what you are doing. Chan et al. (2009) point out that post companies effectively responded to changes, if the labour could perform various tasks. On a day-to-day level they have to handle the varying volume level per day and per week, but also to manage peak periods like the Christmas season. PostNL calculates the expected volume (amount of mail per kind of mail) per day in order to enable process managers and team leaders to manage and to schedule their postmen efficiently. Therefore, PostNL has also defined that the conduct of employees, staff and line has to be in line with the characteristics of a flexible delivery network as part of its strategy. Based on our stakeholder analysis (see Section 3.1) we can confirm the importance of employee commitment for ensuring flexibility. According to the line committed employees tend to be more flexible, to deliver more efficient and to show a better behaviour towards customers.

Quality
Quality means that you want to do things right. PostNL defines good quality as delivering on time to the correct address without damage. Therefore, quality depends on the performance factors speed and dependability, which are, according to Chan et al. (2009), critical for the postal service. Judging the quality performance of PostNL from a market perspective PostNL offers a slightly better performance than its main competitors in the national mail sector Sandd. Sandd delivers 96.0% of the mail on time and without damage to the right address (Sandd, 2016b), whereas PostNL is slightly more reliable with 96.4 % (PostNL, 2016a). Most customers take delivery on time and without damage for granted, but if the quality falls below the standard PostNL might lose customers. Thus, it is critical for PostNL to keep the quality level high. This can also be seen from the bottom-up perspective where line employees of PostNL priorities customer satisfaction in relation to the delivered quality the most.

Even if the priorities of CSFs are different, during the analysis we have seen that most CSFs are interrelated and thus should all be considered in the performance analysis of the mail delivery process. For instance, in order to have efficient postmen it would help to gain their commitment. If postmen are committed they are more willing to work flexible. Flexibility enables to keep a high quality (delivering on time, to every address, without damage) even in case of unexpected changes in mail volume. Using depots and means of transportations more efficiently, increases cost efficiency and enables to offer lower prices to customers and still to keep a reasonable profit. This interrelation between those areas is also shown by Slack et al. (2010) (see Table 5.1) indicating that only if all those factors performed sufficiently, we could work cost efficient. Thus, a company cannot solely focus on costs minimization, but still has to ensure the level of quality, speed, dependability and flexibly that its customer requires. To measure how successful operation is in doing so, we can use productivity as an indicator (Slack et al., 2010).
Currently, the CSF costs, quality, customer satisfaction and employee commitment are already incorporated into the MJ dashboards (see Table I.2), but evaluated independently from each other. Taking productivity as performance measure for the benchmarking provides three main advantages. Firstly, it would solve the current problem of too many performance measures and too low information processing (see Figure 3.1) as it combines all CSFs in one number, which makes it possible that the process managers can directly see if they perform well or not. Secondly, setting the ratio instead of absolute numbers enables a better comparability which is critical for a benchmarking (Neely et al., 1997). Finally, as productivity among others combines the factors costs, flexibility and quality, which are also incorporated in the MJ dashboard, the process manager will see the value and will be motivated to improve it (see Section 3.1). In the following section we will define which

performance measures are required to represent and calculate productivity of mail delivery performance.

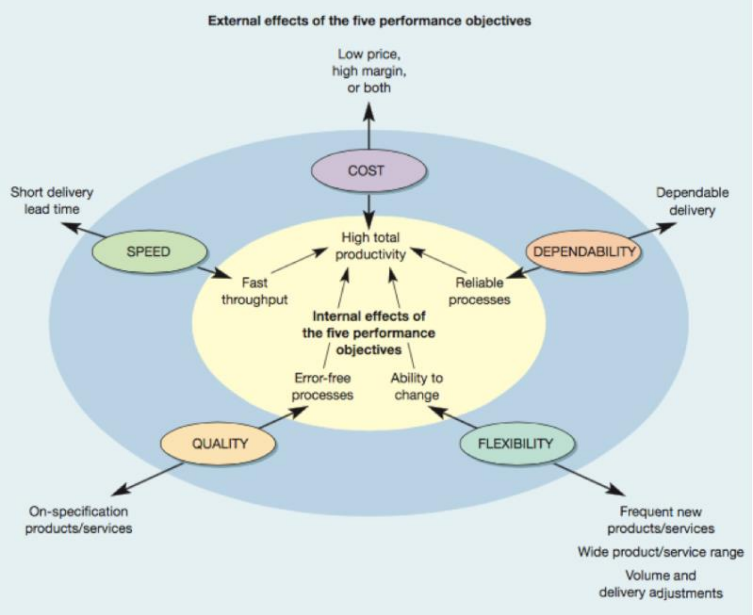


FIGURE 5.2: RELATION BETWEEN THE CRITICAL PERFORMANCE FACTORS (Slack et al., 2010, p. 52)

5.2. From Critical Success Factors to Performance Measures

After having identified the CSFs of mail delivery process and detected that productivity considers all those factors, we have to define what we want to measure in order to represent the productivity of mail delivery process.

Operational productivity is defined as "the ratio of what is produced by an operation or process to what is required to produce it" (Slack et al., 2010, p. 666), given that the quality of the outputs is kept constant (Grönroos & Ojasalo, 2004):

$$Productivity = \frac{Output\ from\ the\ operation}{Input\ to\ the\ operation} \mid Constant\ quality\ of\ outputs$$

The next step is to define the elements input as well as output that we want to measure for the mail delivery process. One framework for identifying concrete measures for a process has been developed by Brown (1996). He emphasizes the difference between input, processing system, outputs, outcome and goal by suggesting that each step has its own specific performance measures (Neely et al., 2000). Based on the mail delivery analysis in Chapter 2 and the four perspectives on operations in Chapter 5.1, we can fill in the framework as follows (see Figure 5.3):

- **Input**: The input of mail delivery is the mail that needs to be delivered, the points that get a delivery and the infrastructure in which the delivery points are embedded.
- **Processing system**: This incorporates the resources from the organisation to support the process. From the operations resource perspective, we can define employees (the postmen), depots and means of transportation as main resources.
- **Output**: The output implies the resulting costs. We can differentiate between delivery costs and depot costs. In contrast to the definition given above, we can also define the delivered quality as a process output, which was highly emphasized by all stakeholders of the benchmarking model.
- **Outcome**: The outcome of the process is the customer satisfaction.
- **Goal**: Based on the top-down analysis of the mail delivery, we can define a goal to create a flexible network and to deliver high quality at low cost.

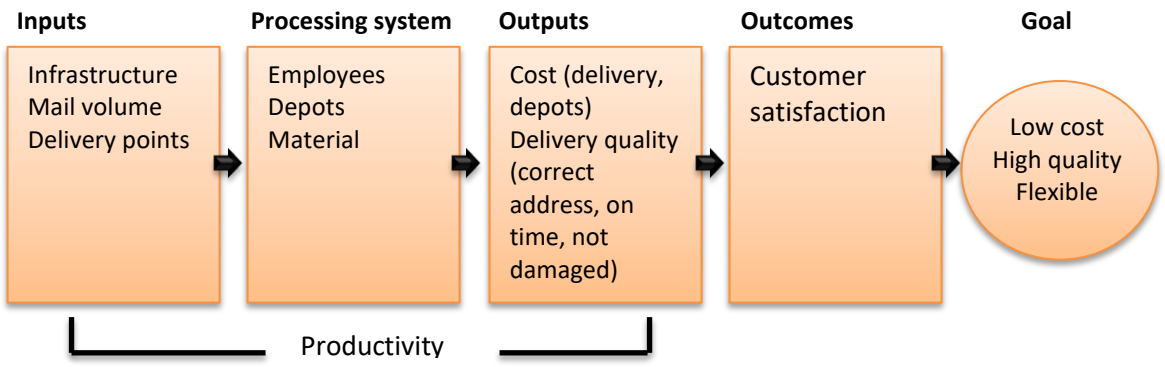


FIGURE 5.3: INPUT, PROCESSING SYSTEM, OUTPUTS, OUTCOME AND GOAL FRAMEWORK OF BROWN (1996, CITED BY Neely et al., 2000) FOR POSTNL

Based on the findings we can see that the operational definition of productivity is not adequate as it does not consider quality as an output. This is mainly because productivity in context of a production process has not the same focus as productivity in context of a service process. In production the focus rather lies on quantity as an output. However, in service productivity cannot be separated from quality (Grönroos & Ojasalo, 2004; Sahay, 2005). We can see this at the mail delivery of PostNL, where it is not only important how many items are distributed per hour, but also how customers receive the service. Thus, measuring productivity for a service one has to consider the quantity as well as the quality.

Furthermore, our analysis shows that flexibility is highly important as well. Grönroos and Ojasalo (2004) point out that demand influences productivity: Low demand leads to underutilisation of the service resources, which does not influence the quality, but it means a decrease of internal productivity. If the demand is higher than what the service resources can handle, resources are fully utilised and hence have high productivity. However, the external demand cannot be met on time, and thus the overall perceived quality is decreasing. Hence, it is critical for the service to have a high flexibility to quickly adapt supply to demand. For manufacturing this is less critical as it can produce inventory to deal with the varying demand. (Grönroos & Ojasalo, 2004)

To sum up, in addition to the traditional measurement of productivity it is highly important to incorporate perceived quality as well as flexibility in terms of quickly responding to demand changes as indicators of productivity. Therefore, we do not follow the definition of operational productivity, but the service productivity model of Grönroos and Ojasalo (2004). They define service productivity in three sub-parts: (1) internal efficiency, which is how efficient input is transformed into output, (2) external efficiency, which is how well quality of the service is perceived, and (3) capacity efficiency, which is how effectively capacity of the service process is utilized considering the demand.

For the mail delivery process of PostNL we can adapt it as follows (see Figure 5.4):

- **External efficiency**:
  As an indirect measure of perceived quality, we can use the customer complaints. During the stakeholder analysis we saw everyone agreeing that complaints are a critical element of performance evaluation, which shows that complaints are highly relevant (Recommendation 2a, Section 4.2). Number and kind of complaints are given per day and per delivery tour. To make it comparable we should use a ratio rather than absolute numbers (Recommendation 3b, Section 4.2), thus, the number of complaints per actual delivery point of a certain time unit. For a more detailed analysis we could also differentiate between the types of complaints, for instance, complaints due to delivering mail to the wrong address, due to damage at the mail or to misbehaviour of a postman.
- **Capacity efficiency**:
  One week in advance the team leaders have to estimate the required delivery time per day given an estimation of the mail volume per day. The realised time can be measured based on

the plus and the minus hours made by postmen. This measurement is relevant to control and determine the budgets for the next quarter. However, it is less relevant for cost efficiency as PostNL pays the same salary for overtime as for normal hours. Furthermore, it always sets the contract hours around 20% lower than the expected working time to ensure that a postman does not gain minus hours.

- **Internal efficiency**:
  We have defined the output of mail delivery process as costs and quality. Quality has already been considered in the external efficiency, and thus costs would be more relevant. At the input we defined three factors; volume, delivery points and infrastructure (see Figure 5.3). The infrastructure is constant so it can be rather be used as cluster attribute. More relevant are volume and number of actual delivery points for the performance measure. Both, volume and actual number, of delivery points are interrelated: the higher the volume, the higher the chance that a delivery point receives mail, which we outline in detail in Section 5.4. The more delivery points a tour has, the higher the volume tends to be. Thus, in the end we have to consider both factors in the measurement, either in the formula for the performance measure or in the clustering. During the clustering we decide how to incorporate those factors exactly.

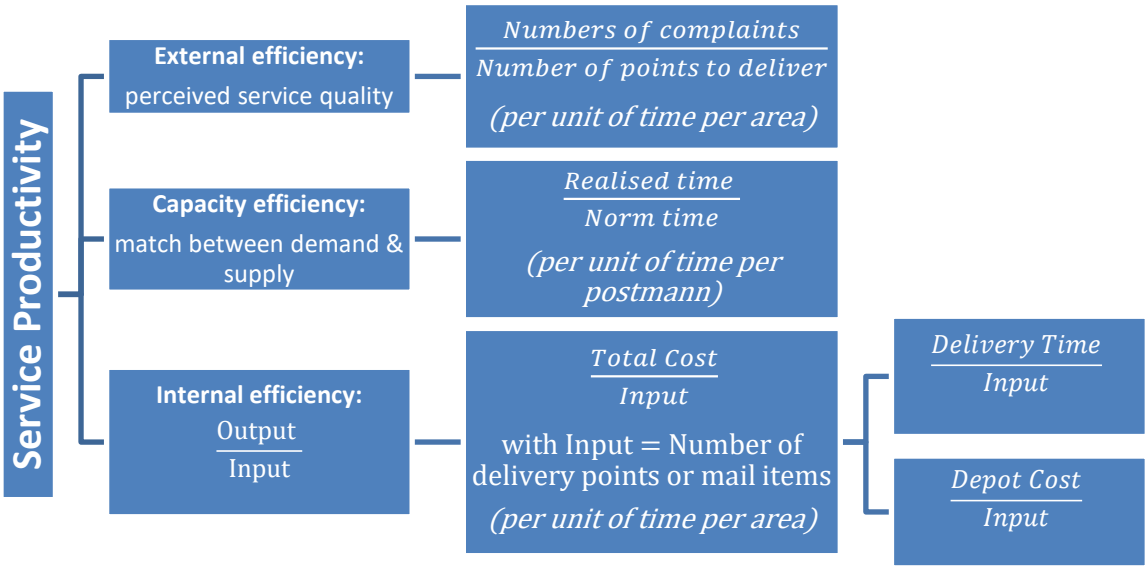


FIGURE 5.4: SERVICE PRODUCTIVITY OF POSTNL

For this research we determine the clustering for only one performance measure according to our scope. Capacity efficiency is less relevant as improving it would not increase cost efficiency. External efficiency is relevant, but as it depends on only two factors, number of complaints and number of actual delivery points, it is less complex than internal efficiency. Thus, if we can develop a clustering technique for internal efficiency, it should be easily applicable to external efficiency. Therefore, we select internal efficiency as performance measure for our clustering.

As mentioned above, internal efficiency depends on costs as input. Within the mail delivery process there are two main cost factors: depot costs and delivery costs. In Figure 5.6 we have clarified the composition of the costs by showing the main factors that impact the overall costs. We can see that depot and delivery costs are influenced by different factors. Thus, for the clustering it would be advisable to give internal productivity two separate performance measures, one for depot and one for delivery costs (see Figure 5.5). As we want to proceed with only one performance measure, we select the one that is more complex and thus more difficult to cluster. Considering the operational resource perspective (see Appendix VIII) we have already clarified that depot costs depend on renting costs of the location, maintenance and handling costs (e.g. cleaning the location or sorting the mail from the car in the shelves), distances from the depot to the starting points of the different

delivery tours and transport distance from the HUB to the depot. The last factor is out of our scope, but critical for evaluating depot efficiency.
Delivery costs within PostNL are measured in time as the delivery is a service with the only expense on wages of the postman per time unit. If we consider the influence factors of delivery time (see Figure 5.6), we can distinct between distances that the postman has to cover (distance between depot and delivery tour, length of main, minor and connection route, distance to retailer) and means of transportation which define the speed. The only factor that is independent of distance and means of transportation is the time to drop the mail into the mailbox. To enable an accurate elaboration and due to the time restriction we set the focus on delivery costs as input for internal efficiency for our research. Delivery costs are not only a bigger component of the total costs but also due to their many influence factors a more complex one than depot costs. Thus more value can be created by developing a clustering technique for the performance measure with delivery cost as input.

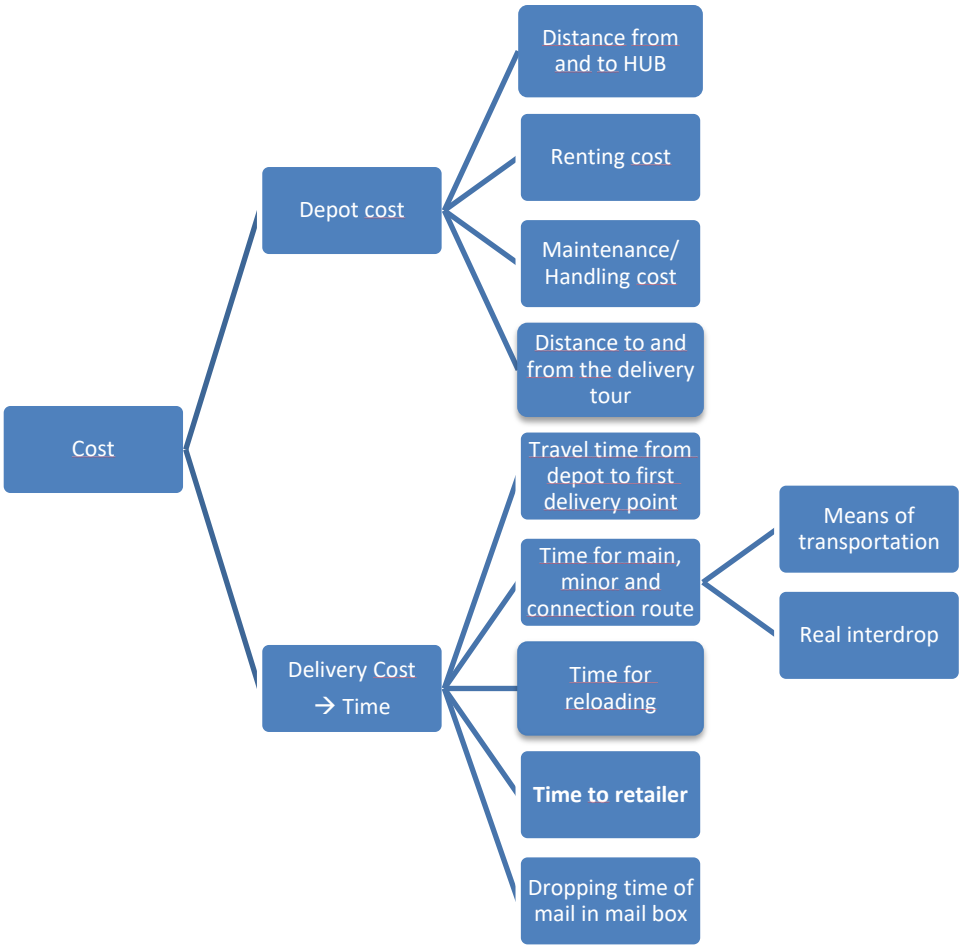


FIGURE 5.5: COMPOSITION OF THE COSTS OF THE MAIL DELIVERY PROCESS

5.3. From Performance Measure to Cluster Attributes

The aim of this section is to limit the number of potential cluster attributes to only those that have a high impact on delivery time and thus on the delivery costs. Therefore, we use three different approaches and derive a list of attributes by comparing the outcomes. The first approach is based on the findings of the stakeholder analysis, where we ask postmen, team leaders, process mangers and staff members to name the most important influence factors for mail delivery. The second approach is based on the time-activity model of the costing and economics department, where the average expected time is calculated for each activity of the mail delivery process. Finally, based on already existing norm models for calculating the delivery time, we assess the most critical attributes together with an expert team.

Stakeholder Analysis

As stated above we have conducted semi-structured interview with 12 different stakeholders of the mail delivery process. Into this interview we have also included a question on the factors that might influence the mail delivery performance. The most frequently named attributes are shown in Table 5.2. The overall list of attributes can be found in Appendix V.

TABLE 5.2: MOST IMPORTANT ATTRIBUTES BASED ON THE STAKEHOLDER ANALYSIS

Attribute	Frequency
Location of the mailbox	5
Interdrop	5
Address arrangement	4
Number of delivery points	3
Kind of mail	3
Customers	3

One of the most frequent named aspects is the layout of houses, in particular because it has influence on two attributes: firstly the location of the mailbox and secondly the interdrop. For instance, the mail delivery to a house with a front yard takes longer than delivering mail to flat buildings with one front of mail-boxes, because the minor-route that needs to be passed in order to access the mailbox is much higher given the yard. Furthermore, flats often have one front with the mailboxes for all residents, thus more mail can be delivered in less time due to the small interdrop and thus high density of mailboxes. Moreover, the location of the mailbox also includes the aspect on which floor it is located as it is much more time-consuming to take stairs compared to walking on a flat path.

The third most named aspects, which is in particular named by postmen, is the address arrangement within a street; sometimes their mail packages are ordered in a sequence different to the addresses. However, this is rather an indication that the data-base of PostNL is not on order as PostNL aims to sort the mail according to the address arrangement. Therefore, we are not including this in our clustering.

Fourthly, the number of delivery points influences the delivery time as for each delivery point the postman has to select and grab the right mail and put it into the mailbox.

Finally, the kind of customer can influence the time, in particular due to special wishes on the deliver location (in the mailbox, at the reception). However, officially every address in the Netherlands has to have a mailbox outside its building according to the postal law and PostNL expects from postmen to deliver mail always to the mailbox except if it does not fit or if the customer agreed with PostNL on special terms. Therefore, we will not consider this aspect as a cluster attribute.

Overall, based on the stakeholder analysis we select the following as potential clustering attributes: minor-route (stairs, flat), interdrop and number of delivery points.

Activity-Time Model

The costing and economics department has designed a model showing the impact of an activity on mail delivery time. Therefore it determined the processing unit and expected time per processing unit for each activity. To calculate the impact, the number of units per day is estimated and multiplied by the expected time. This model is updated at least once a year and gives the average for all areas within the Netherlands. Furthermore, the model makes a distinction between the means of transportation (walk/bike, scooter and car) and weekday. They also differentiate between the kinds of mail. For instance small mail items can directly be put into the mailbox, while all ring packages, which are packages that require a signature or do not fit through the mailbox, need the activity “ring and wait at the door”. In the following we will determine the most time consuming activities for mail delivery and parameters influencing those activities for each means of transportation. Subsequently

we will analyse the impact of the kind of mail on the mail delivery process to determine if it is necessary to differentiate the kind of mail in our benchmarking model.

TABLE 5.3: MAIN ACTIVITIES AND THEIR % OF TIME CONSUMPTION PER MEANS OF TRANSPORTATION (B/F: BIKE/WALK, S: SCOOTER, C: CAR)

Activity	Processing unit	Parameter	Tuesday			Wednesday			Thursday			Friday			Saturday		
			B/F	S	C	B/F	S	C	B/F	S	C	B/F	S	C	B/F	S	C
Run-up	Delivery tour	Length of run up	3%	2%	6%	3%	2%	6%	3%	2%	6%	3%	2%	6%	4%	2%	7%
Biking/ walking/ driving main-route	Real Interdrop	Length of the interdrop	55%	74%	74%	68%	77%	77%	54%	74%	74%	69%	78%	78%	57%	80%	80%
Walking minor-route	Delivery points to be visited	Average length of the minor-route	17%	9%	7%	11%	8%	5%	17%	10%	7%	11%	8%	4%	15%	7%	7%
Deliver at delivery point	Delivery points to be visited	Time for putting mail in mailbox	21%	12%	33%	15%	10%	20%	22%	12%	34%	14%	10%	20%	20%	9%	29%
Riding run-off	Delivery tour	Length of run off	-	-	6%	-	-	6%	-	-	6%	-	-	6%	-	-	7%

In Table 5.3 we present the activities that consume more than 1% of the delivery time per tour together with their processing units and parameters. We can see that, independent of the means of transportation, the activities are the same, except the run-off activity for car-deliveries. This is mainly due to the fact that bike and scooter deliveries do not have to return to the depot anymore. Furthermore, even if the activities are the same, the percentage of time consumption can differ between them. For all means of transportation we can state that covering the main-route followed by delivering mail at the delivery point consumes most time. Covering the minor-route and run-up of a tour have less impact.
Considering this model, we can identify the length of minor-route, interdrop and run-up as potential cluster attributes as well as the number of delivery points to visit and the time for putting mail in the mailbox. Furthermore, based on this model we can see an impact of peak- and off-peak days on the time for each activity. Looking at the exact calculations of the model, we can see that some factors are volume dependent (such as the number of delivery points and the actual interdrop length) while others are constant (expected minor-route per delivery point, time to put mail in mailbox, run-up distance). This can have an impact on our clustering, which we will analyse later on in Chapter 6.

This model helps us to assess the impact of the different kind of mail on delivery time as it can calculate the expected time based on the probability that a mail sort requires certain activities and the time of the activities. Again we have differentiated between means of transportation as not all activities are the same (for instance, deliveries by car do not have to go to the retailer, but can return the not delivered mail to the HUB).

TABLE 5.4: RELATIVE WEIGHTING FACTOR PER ITEM GIVEN THEIR COST

Kind of Mail	Bike/Walk	Scooter	Car	Average
Small	1.0	1.0	1.0	1.0
Big	1.0	1.0	1.0	1.0
Mailbox packages	2.4	1.6	1.9	2.0
Ring packages	4.3	2.5	2.7	3.2

Table 5.4 shows the relative weighting factor per kind of mail based on the average cost per item relatively to the small mail. We can see that there is a significant difference between bike/walk deliveries compared to scooter and car, which is mainly due to the time it takes to go to the retailer.

Furthermore, for each means of transportation we can state, that ring packages weight much more than any other kind of mail. The weight difference can be explained based on two activities and the probability that this activity is required given a certain kind of mail: firstly the activity "ringing and waiting at their door", which takes around 0.5 min and is not required for small and big mail as it fits in the mailbox. Around 8% of all mailbox packages do not fit, and thus postmen have to ring and wait. 100% of ring packages have to follow this activity. The second activity is "going to the retailer", which is highly time consuming and only required if the item cannot be placed at the neighbours' house. Overall, this activity is required for around 0.24% of the mailbox packages and 3.04% of the ring-packages.

Overall, the model of costing and economics shows that the impact of factors highly varies per means of transportation. Therefore, we should consider means of transportation as a cluster attribute. Furthermore, impact of each the kind of mail is highly varying. Thus, if we want to calculate the time per mail item as a performance measure, we should give each kind of item a different weight or ensure that in each cluster the proportion of the sort mail is the same.

Norm Model

The logistic strategy department has designed a model for calculating the exact delivery time per day and applied it on twelve delivery tours. Together with a group of expertise (process manager of optimisation, senior manager of logistic strategy, senior controller) we have selected the elements with the highest influence on delivery time. Subsequently, we have eliminated those with a low occurrence among all delivery tours and finally determined parameters to calculate those elements.

1. **elements of the delivery time**
   Delivery time = time for run-up + main-route time + minor-route time + serving time + time to retailer
2. **eliminating elements with low occurrence**
   The expert team has decided to define critical elements as those that have a high time consumption as well as a high occurrence. For instance, the time to the retailer is generally high, however it is often not necessary for postmen to go to the retailer as it is mostly possible to deliver mail at the delivery point or to pass it to neighbours. Furthermore, since 2017 the run-off has not been within the budget anymore, because postmen do not have to go back to the depot. However, an exception are deliveries by car, which are less than 4%, as cars have to be return to the HUB at the end of the day. Thus, the time to retailer and the run-off time can be ignored. However, the experts emphasised that if we divide delivery time by the number of mail, we should give ring-packages a higher weight as on average the occurrence might be low, but per area the number of packages can highly differ and thus should be incorporated to some degree in the performance measure.
   Finally, the run-up time is independent booked in the contract time for each delivery tour at PostNL as the run-up time is independent of the delivery area, but also depending on the location of the depot. Thus the length of the run-up can rather be incorporated when benchmarking the depot structure rather than the mail delivery process.
3. **calculating the elements**

$$Main - route\ time = \text{mainroute length} * V_T$$

$$Minor - route\ time = \frac{(\sum Minor\ route\ \ to\ cover\ )}{5\,\frac{km}{h}}$$

$$Serving\ time = \ APN\ to\ visit * normtime\ for\ serving$$

$$with\ V_T = velocity\ of\ means\ of\ transportation\ T$$

Based on the formula to calculate each element, we can see that the main aspects which are important are the distance the length of the main- and minor-route as well as the number of points to deliver (APN). The key factor of determining the time of the main-route is the means of transportation as they determine the speed to cover the distance. As mentioned above, means of

transportation depends on the length of the interdrop, thus clustering on the interdrop automatically ensures that the means of transportation is the same. For the minor-route it is less critical as most of the time, independent of the means of transportation, it is walked. Within the norms for calculating the minor-route, there is a differentiation between flat and stairs. One meter of stairs is weighted equal to 1.4 m of flat distance to account the higher time consumption of taking stairs.

Comparison and Conclusion

The attributes derived from the different approach have a high overlap. Each approach derives with interdrop, number of delivery points to visit and minor-route as attribute. Two of the three approaches emphasise the difference between stairs and flat minor-routes and thus should also be considered when determining and measuring the final cluster attributes.

Furthermore both models differentiate between the means of transportation, due to the different possible speed, but also because they incorporate different activities (e.g., going to the retailer, run-off). However, we can see based on the frequency of visiting the retailer, that this is less relevant for foot, bike and scooter deliveries and can thus be ignored as cluster attribute. The run-off has quite an impact on car deliveries, but because less than 3.5% of delivery points are delivered by car, it is irrelevant for most delivery areas and thus can be excluded for the clustering. Considering that the means of transportation is selected based on the interdrop, we assume that if the interdrop within a cluster is similar, the means of transportation should be as well.

Moreover, the costing and economics as well as the expert-norm model approach indicated that each kind of mail has a different impact on the mail delivery time, however the frequency considering all tours in the Netherlands is low and thus less relevant as a cluster attribute. Instead it should be incorporated into the performance measure by giving each kind of mail a different weighting factor. Therefore, the internal performance measure should be the delivery time divided by the weighted number of mail items. For the weighting factor we can use the weight suggested by the costing and economics department or if desired conduct a separate study to define them.

Another similarity is that both models incorporated the main-route and directly link it to the interdrop. Looking at the norming model, we can see that the interdrop is especially important to determine the velocity for covering the main-route.

Finally, based on the costing and economics model, the run-up distance should be considered as possible cluster attribute, but has a quite low impact with 2% to 6%. Referring to the experts it is also a less relevant aspect of the mail delivery process, but rather more important for evaluating the depot infrastructure.

Thus, overall possible cluster attributes are interdrop, especially for determining the means of transportation, delivery points to visit, minor-route and main-route distance.

5.4. Information Requirement, Availability and Validation

Before proceeding with the final selection, we have to ensure that we are able to measure the attributes. Therefore we will discuss the information availability on kind of mail, interdrop and the connection to the means of transportation, APN, minor-route and main-route distance in the following. In case of an estimator for the attributes, we will assess their accuracy and possibilities for improvement. Finally we will discuss the cluster objects for our cluster analysis, which depends on the level on which information of the performance measure as well as the cluster attributes is available.

Number of Delivery Points

PostNL knows the exact number and addresses of the delivery points (APN) for the Netherlands and stores its data base, “Base Register of PostNL” (BRPP), which is regularly reconciled with the Base Register of Addresses and Buildings (BAG) of the Dutch Government. In the near future (2018) all

sorting machines of PostNL can provide exact information on the number of mail items per address (PostNL, 2016c) and by that the exact number of houses that receives a delivery. Until then PostNL estimates the number of houses based on the hit-chance (HC), which is the chance that a house receives mail in area u.

$$HC_u = 1 - \ exp^{-\left(\frac{v_u}{\mathrm{APN}_u}\right)} \tag{27}$$

This estimator depends on the volume (v) and number of delivery points (APN) of a given area u. Overall, the expected number of points within area u ($\widehat{\mathrm{APN}}_u$) that actually receive a delivery can be calculated as follows:

$$\widehat{\mathrm{APN}}_u = \mathrm{APN}_u * \ HC_u \tag{28}$$

A study conducted in 2016 by PostNL showed that this estimator overestimate the number of delivery points by 3.11% on postcode level 6 and 5.47% on delivery tour level. This level of accuracy is acceptable for our study.

Minor-Route

The minor-route (see Figure 2.1 – No.5) of each delivery point is measured manually with a surveyor's wheel. PostNL distinguishes between flat and stairs distances and stores the information in the data base BRPP. We assume a high accuracy for the flat part ($dimi_{flat}$) of the minor-route as PostNL sells this information to other organisations and thus sets great value on accuracy. However, the distance calculated for stairs ($dimi_{stairs}$) is less reliable due to two reasons. Firstly, because walking stairs takes longer than flat distances, stair distance is supposed to be equalised to the flat distance by multiplying each meter of stairs by 1.4 within BRPP. However, the norm 1.4 was defined 40 years ago and never revised and thus might be outdated. Secondly, PostNL does not resell it, and therefore accuracy is less emphasised. Therefore, this data should be handled with caution.

As we do not know the exact delivery point that receives mail, we use the average minor-route per delivery point (dimi) for each cluster object u:

$$dimi_{\ u} = \frac{\sum_{a \in u} dimi_a}{APN_u} \tag{29}$$

with $dimi_a = \ dimi_{\ a,flat} + dimi_{\ a,stairs}$

and a = delivery point given in BRPP

Main-Route

The length of the main-route (see Figure 2.1 – No.3) is estimated with the GPS network and street system Geodan. Geodan allocates each coordinate of a delivery point to a coordinate on the street and measures the total distance of covering all coordinates on the street given the tour sequence. Consulting the route designer of PostNL, we can assume that the distance measure is quite accurate. However, the drawback of Geodan is that in rare cases it connects the delivery point to the wrong street, leading to a less reliable measure. PostNL tries to increase the reliability by rechecking the assignment during the design stage of a tour. Therefore, we assume that the accuracy of this estimator is sufficient for our study.

$$dima_u = \sum_{i \in u} d_{i,j} \tag{30}$$

Interdrop

There are three approaches on calculating the expected distance between delivery points, the interdrop, given the delivery points (APN_u). As mentioned above the delivery points to visit are approximated by multiplying the APN_u with the hit-chance (HC_u) of a given cluster object u.

The first approach is based on the surface of the cluster object (S_u) and the number of delivery points within that surface.

$$expected\ interdrop_1 = \sqrt{\frac{S_u}{\mathrm{APN}_u * \ HC_u}} \tag{31}$$

This measure can be used if no information is given on the delivery sequence of the delivery points. However, it assumes that delivery points are uniformly distributed on the surface, which is often not the case. Thus it is an easy way to calculate this estimator, however not highly accurate.

The second approach is based on selecting randomly $\widehat{\mathrm{APN}}_u$ delivery points within cluster object u and calculating the linear distance between their GPS coordinates of the delivery points (i,j) $(dG_{i,j})$ given their sequence of visiting.

$$expected\ interdrop_2 = \frac{\sum dG_{i,j}}{APN_u * HC_u} \tag{32}$$

This indicator would be already more accurate than the expected interdrop$_1$, but requires more computational time and the delivery sequence.

Thirdly, the expected distance can be based on the road network system, which is also currently used by PostNL. Based on the geo-information provided by the company Geodan, PostNL can estimate the distance $(d_{i,j})$ from one delivery point (i) to the next following delivery point (j) given a certain delivery tour. This distance is measured in meter. For the accuracy accounts the same as for the accuracy of the main-route distance, which is assumed to be high. In order to calculate the expected interdrop, PostNL uses the following estimator:

$$expected\ interdrop_3 = \frac{\sum_{i \in u} d_{i,j}}{APN_u * HC_u} \tag{33}$$

By that we take the sum of distances between all delivery points (i) and its succeeding delivery point (j), independent if j is within cluster object u or not. Thus, we include those distances leaving the cluster object, but not the distances from a delivery point outside the cluster object to a delivery point within the object. The drawback of this estimator is that it includes the connection route, which results in an overestimation of the interdrop.
To improve the interdrop estimator we suggest to subtract the distances which are part of the connection route:

$$expected\ interdrop_4 = \frac{\sum_{i \in u} d_{i,j} - \sum_{d_{i,j} \in CR, i \in u} d_{i,j}}{APN_u * HC_u}. \tag{34}$$

The connection routes (CR) are defined by PostNL as the part of the main-route where no delivery takes place, but only travelling from one area of delivery points to another area of delivery points. Currently, there is no direct information given on the connection route distance at PostNL. However, each group of delivery points is stored in the information system of PostNL and defined as one delivery section. Furthermore, information on the exact delivery sequence is given and the distances between each delivery point to the subsequent delivery point as well. Thus, by extracting the interdrop between the last of a section and the first delivery point of the succeeding section, we can obtain the distance of the connection routes. However, sometimes groups of delivery points can also be directly next to each other (for instance two flats with a distance of 5m). To ensure that the distance between delivery groups is sufficiently large we set a minimum of 15m in order to define a distance as connection route.

Assuming that the expected interdrop$_4$ is the most accurate, we compare it to the expected interdrop$_1$ and interdrop$_3$ based on the sample of delivery area Utrecht with around 326000 delivery points. The expected interdrop$_3$ is not considered as it provides no advantage on the computation time nor on the accuracy.

On average the expected interdrop$_1$ is 5 times higher than the expected interdrop$_4$ with an average absolute bias of 19m. Therefore, even if the expected interdrop$_1$ is easy to calculate, it highly overestimates the interdrop and therefore should not be used. Finally, given that the expected interdrop$_3$ is 2 times higher with an absolute difference of 10m, we conclude that the slightly higher

computational effort for the expected interdrop$_4$ is adequate. Thus overall, we proceed with the expected interdrop$_4$ for our further research.

Means of Transportation

PostNL has developed a model for selecting the optimal means of transportation. This model selects the means of transportation that is the cheapest given a certain interdrop, which depends on the velocity of the means of transportation and the cost per hour for a means of transportation. This model defines exact turning points for the means of transportation given the interdrop. However, in this paper, due to confidentially, we only name the range in which a turning point lies (see Table 5.5).

TABLE 5.5: TURNING POINTS FOR THE MEANS OF TRANSPORTATION GIVEN THE INTERDROP

Means of transportation	Interdrop in m
foot/bike	0
bike	20-40
e-bike	40-60
scooter	50-70
car	100-120

For instance, delivery by car is optimal given an interdrop higher than 100-120m (see Table 5.5).

Kind of Mail

PostNL knows the exact amount of mail items for each delivery tour and also for each postcode level, stored in the Network Volume Registration (NVR). However, we are limited when determining the volume per kind of mail. Until now the machines are only able to differentiate between small, big mail as well as packages, but cannot differentiate between ring and mailbox packages. However, in 2018 all sorting machines of PostNL can provide the exact number for each kind of mail. Until then, we use the number of all packages as an estimator for ring and mailbox packages combined. In consultation with the volume coordinator of PostNL we assume that the proportion of mailbox and ring packages is the same in all areas. Therefore we suggest to use for the weighting the average between the weight for mailbox and ring-packages.

Cluster Object

In order to test the distribution and correlation of the attributes, we first need to define the cluster object or in other words the areas in which we measure the number of APN, the interdrop, the minor-route and the major-route. Overall we can choose cluster objects based on the organisational structure of PostNL (delivery area, team or delivery tour), the postcode level (postcode 6, postcode 5 or postcode 4) or design our own geographical grid.

Taking the organisational structure as clustering gives the advantage that it is easy to implement for the benchmarking. Furthermore, for our performance measure - realised delivery time per weighted mail item - we require the realised time as input, which is only available per team and per postman. Thus, from that perspective clustering on team level is the most advisable. However, one team consists of 30 to 60 delivery tours, which might be too large and imprecise to identify the correlation and distribution of cluster attributes when taking the average of all those tours. Another drawback is that the true cluster structure might not be identified, because as mentioned at the beginning of this report, the organisational structure is build based on the proximity of delivery areas, but independent on the characteristics of an area, therefore using it for assessing the relation between attributes might be misleading. Finally, PostNL is currently restructuring the delivery tours, thus the clustering would not be useful for PostNL in the future.

Another approach is to use the postcode as cluster object. Figure 5.6 shows the setup of the Dutch postcode, where we can see that one postcode gives information about different levels of areas (region, district, neighbourhood or street). For our clustering we have to ensure that the cluster object is homogeneous, thus the smaller the area, the higher the chances of homogeneity. However, to assess the relationship between our attributes, we have to ensure a sufficient size. If the size is

too small, some estimations are less accurate or cannot be calculated, for instance, an area with only one delivery point does not give any information on the expected interdrop.

Considering the delivery area of Utrecht (Postcode 3400-4000), PC6 is the smallest and refers to geographical areas like streets (including the adjacent land and buildings) with an average area of 19000 m² and 18 delivery points. On the other hand PC5 refers to neighbourhoods with an average area of 284000m² with 270 delivery points. Finally, PC4 level refers to districts with an average area of 5km² and 2800 delivery points. Given this, we conclude that PC6 level is too small as an unit of measurement, while PC4 is with 5km² quite large and thus the risk of heterogenic areas is quite high. Therefore, when using the postcode as cluster object, we will use PC5. Drawbacks are that the area size of each the postcode varies highly independently of the postcode level. Given PC5, we have a standard deviation of 706000m² with a mean of 284000m², indicating high variation in the size which might have an impact when analysing the relationship between delivery points and the interdrop.

FIGURE 5.6: SETUP OF A POSTCODE IN THE NETHERLANDS

A third approach is to design a grid for the Netherlands (see Figure 5.7), which ensures that each cluster object has the same area size. Furthermore, this approach is independent of any current management or postcode system, providing two advantages: firstly, the clustering based on this approach is still valid, even, if the management or postcode system changes, which can be expected based on the developments within PostNL. Secondly, the size of the area can be determined by us, thus we can select most optimal area size for our cluster objects. The drawbacks are that PostNL measures mostly data on postcode or delivery tour level. Thus, to analyse cluster attributes we first have to recompose the data. Whereas we can easily determine the number of delivery points per square as we have the GPS coordinates given, this is currently not precisely possible for the mail volume as the information is only given per tour and per postcode level. There are different possibilities to estimate the volume, for instance based on the percentage of PC4 within a certain square. However, calculating this estimator and recomposing the data of the other cluster attributes is highly time consuming and thus not the most practical approach.

FIGURE 5.7: GRID APPROACH

Overall, the grid approach would be optimal from a theoretical perspective. However, the current information infrastructure cannot be easily transferred to it yet. We expect that it will be possible in the near future, when all information is available per address. The postcode approach would be less optimal, but easier to apply. To minimise the difference due to different area sizes of a PC5 area, we can adjust our cluster attributes. In particular, we take instead of the expected number of delivery points, the number of delivery points per km² of cluster PC5 area u as cluster attribute for cluster object PC5. The area size of a PC5 can be extracted from Geodan. Furthermore, as the length of the main-route also depends on the area, sorely the main-route might not show patterns with the other attributes. Therefore, we test next to the main-route also the number of delivery points per meter main-route as cluster attribute (APN/m).

The organisational structure approach would be neither advisable form a theoretical perspective nor from a practical perspective, given that PostNL will change the delivery tours. Therefore, we use the postcode approach for our cluster, in particular the PC5 areas are the cluster objects.

The final pre-selected cluster attributes given PC5 areas as cluster object u are shown in Table 5.6 including their measurement unit, source of information (Geodan, basis register of PostNL (BRPP), the Network Volume Registration (NVR)) and their formula. For simplicity we refer to the expected number of delivery points and the expected interdrop sorely as number of delivery points and interdrop from now on.

TABLE 5.7: FORMULAS FOR THE PRE-SELECTED CLUSTER ATTRIBUTES GIVEN PC5AREA (U)

Pre-selected Cluster attributes	Measurement unit	Source of Information (date of data extraction)	Formula
Main-route (dima)	m	Geodan (13-01-17)	$\sum_{i \in u} d_{i,j}$ (30)
Delivery points per meter of main-route	m	BRPP and Geodan (13-01-17)	$\frac{\mathrm{APN}_u * HC_u}{\sum_{i \in u} d_{i,j}}$ (35)
Minor-route (dimi) per APN	m	BRPP (13-01-17)	$\frac{\sum_{a \in u} dimi_a}{APN_u}$ (29) $with\ demi_a = demi_{flat} + demi_{,stairs}$
Delivery points per km²	#	BRPP and Geodan (13-01-17)	$\frac{\mathrm{APN}_u * HC_u}{km^2{}_u}$ (36)
Hit-chance (HC)	%	NVR and BRPP (13-01-17)	$1 - exp^{-\left(\frac{v_u}{\mathrm{APN}_u}\right)}$ (27)
Interdrop (in)	m	Geodan (13-01-17)	$\frac{\sum_{i \in u} d_{i,j} - \sum_{d \in CR, i \in u} d_{i,j}}{\mathrm{APN}_u * HC_u}$ (34)
Mail volume (v)	#	NVR (13-01-17)	Exact number

5.5. Conclusion

In this chapter we answer research question 4 by defining suitable performance measures for the national mail delivery at PostNL and possible cluster attributes. The success of the mail delivery process of PostNL depends on multiple interdependent factors (CSF), including employee commitment and flexibility, cost efficiency and delivery quality. While the current MJ framework measures them independently (see Table 2.1), we propose to use performance measure productivity for the benchmarking model. It incorporates the CSFs and by that enables process managers to see by one number how well they perform, which simplifies the benchmarking. As mail delivery is a service, productivity should be measured internally as well as externally. While internal productivity is the traditional ratio between input and output, external productivity focuses on customer satisfaction, which is also essential for a service provider like PostNL. Within the internal productivity we distinguish between delivery time per mail item and depot cost per mail item as they are influenced by different parameters and thus require different clusterings. Overall, delivery time per mail item has the most parameters and thus it is the most complex of all performance measures. In Section 5.3 we show how to reduce the number of possible cluster attributes by evaluating the impact and frequency of each parameter. Interdrop, delivery points to visit, minor-route and main-route distance are parameters which have the highest impact on delivery time and are relevant in all tours (high frequency). In contrast, the kind of mail has also an impact, especially ring-packages, however, those are less frequent within a tour. Therefore, the parameter mail type is not incorporated as cluster attribute, but instead in the performance measure by giving each kind of mail a different weight. Overall, the final performance measure used for clustering is the delivery

time divided by the weighted number of mail items with the possible cluster attributes interdrop, number of delivery points (APN), minor-route and main-route distance.
For defining the cluster objects we make use of the postcode system. The postcode system is a stable system (no significant change within 40 years) within PostNL and most information is given on a postcode level. Consequently, we can easily gain all the required information on our cluster attributes per cluster object and ensure that the clustering is also applicable in the long-term. In particular we select postcode 5 (PC 5) areas as cluster objects. To minimise the difference due to the different area sizes of a PC5 area we use APN/km² instead of only the APN as a possible cluster attribute and also test APN per meter main-route.

6. Cluster Analysis for PostNL

In this chapter we will conduct a cluster analysis for the mail delivery process of PostNL to determine a suitable clustering for the benchmarking model. In the previous section we have defined internal productivity, in particular the realised delivery time divided by the weighted number of mail items, as our performance measure with the potential cluster attributes interdrop, APN/km², APN/m, minor-route and main-route distance and PC5 areas as cluster object (see Section 5.4).
In the following we will apply the steps of cluster analysis to the mail delivery process, therefore we will first define the test framework in Section 6.1, based on methods derived from the literature review in Section 4.3. The results of the test will be discussed in Section 6.2 with an evaluation and conclusion in Section 6.2.4. Finally we provide an overall conclusion of this cluster analysis in Section 6.3.

6.1. Test Framework

In this section we will present the test framework. It is an adaption of methods and techniques defined in the literature review to the case of our benchmarking model given the performance measure and our knowledge on the preselected attributes. Before that we will define our test sample.

Sample

As data collection and computation time of conducting a cluster analysis for the whole Netherlands would take a long time, we conduct a cluster analysis on a representative sample. Therefore, we select the delivery area (BG) Utrecht, which incorporates the PC5 areas 3400A-3739N and 3989N-3999W with overall 1265 PC5 areas and 326000 delivery points. From an organisational perspective it consists out of 39 teams and 1126 delivery tours with 33 deliveries by scooter, 9 by car and the remaining by bike/walking. Considering the address density, BG Utrecht contains every kind of type, from highly rural to highly urban. Information for calculating the pre-selected cluster attributes per PC5 area are extracted from the databases Base Register of PostNL (BRPP), the geo-network system Geodan and the Network Volume Registration (NVR) of PostNL.

Overall, with the sample size of 1126 cluster objects and the variety of areas, we assume that BG Utrecht is a representative sample. The drawback of taking a sample is that it might not represent the true densities for the whole Netherlands given a certain data space. Therefore, we will not use density based cluster algorithms in our test framework.

Test Framework

As defined within the literature review, cluster analysis consists of four steps which also determine the structure of our test framework; the first part is to analyse and assess the distribution and relations of the attributes. In the second part we apply different weighting methods to the attributes. Thirdly, clusters are formed by applying different clustering techniques, which are subsequently evaluated in the final step. However, before starting the data set need to be prepared.

Preparation: Data collection and Examination

The first step is to select the required data for the attributes. The data has to be given for each object within our sample. As many algorithms are vulnerable to noise and outliers within the data, we examine the data before starting the cluster analysis based on two approaches.
Outliers can be defined as points “that stand away from the body of the distribution” (De Veaux, Velleman, & Bock, 2012, p. 50). Therefore, our first approach is to assess the boxplot for each attribute (see Figure 6.1). Every point above the upper or lower fence can be seen as suspected outlier. With the upper fence = 3rd quartile + 1.5 *(third-first quartile) and lower fence = 1st quartile - 1.5 *(third-first quartile) (De Veaux et al., 2012).

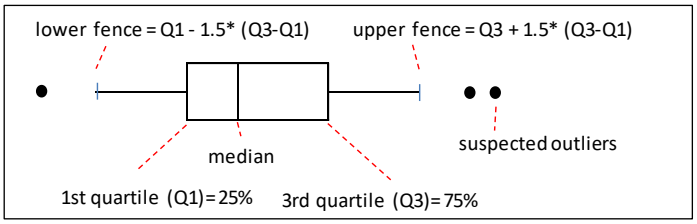


FIGURE 6.1: BOXPLOT WITH UPPER AND LOWER FENCE

The second approach, that we use is to pre-run the data set with the single-link hierarchical cluster algorithm and identify potential outliers based on the distance at which objects are merged, which is shown in the agglomerative table (Ketchen & Shook, 1996). In contrast to the histogram, this approach considers all attributes of a cluster object together.
The decision to eliminate outliers depends on the purpose (Ketchen & Shook, 1996). For clustering, we want to eliminate outliers which are due to the wrong data input. Only if we would analyse the data of the whole Netherlands, we could determine with certainty if a point is an outlier. However, given that we take a sample, it might be the case that the outlier is just underrepresented within the BG Utrecht, thus in that case we keep them. We analyse if the wrong data input is given by assessing with the geo-map if the data of that outlier is reasonable or not. In latter case, we try to find the accurate data or otherwise erase it.

I. collecting data of all possible cluster attributes for each cluster object (data source: BRPP, NVR and Geodan)
II. examining the data set on outliers and noise by analysing the boxplot, using single-link hierarchical clustering and assessing the correctness of their data input. Erase them if necessary.

Attribute Relations and Distribution

Cluster algorithm find and create almost always clusters even if no clusters exist within the data (Tan et al., 2005b; Xu & Wunsch, 2005). To improve the classification performance, it is critical to select only the most discriminatory attributes (Jain, Murty, et al., 2000). Therefore the first step is to assess each potential attribute and their relation with each other.

PostNL differentiates between peak- and off-peak days when estimating the interdrop and the number of delivery points. Reason is the difference in mail volume, which they incorporate by the hit-chance (see formulas in Table 5.7). In contrast, main-route and expected minor-route per delivery point are independent from hit-chance and thus the same during peak and off-peak day. The aim of evaluating the relationship between attributes is to identify intrinsic clusters (cluster tendencies) within the data set (see Section 4.3.2). Due to the difference in volume, there might be a different relation between attributes during peak- and off-peak days. Furthermore, there is the risk that incorporating the volume might blur cluster tendencies between interdrop, delivery points, minor-route and main-route, because while the volume depends on the inhabitants, the pre-selected attributes mainly depend on the infrastructure. Resulting in the following two hypotheses:

H1: The volume included by the hit-chance in estimating the attributes does not contribute to the intrinsic clusters as it is independent from the infrastructure.

H2: The grouping of objects between peak and off-peak is not the same due to the volume difference.

In order to test these hypotheses, we distinguish between three scenarios. The first scenario is only based on the infrastructure (Scenario I), where all delivery points receive mail, thus for all formulas given in Table 5.7 we exclude the hit-chance (HC). The second assuming a peak day (Scenario P), where we use APN_u as defined in Formula 28 with HC given the average volume of a peak day. The third assuming an off-peak day (Scenario O), where we also use APN_u as defined in Formula 28 but given the average volume on an off-peak day. For each scenario we evaluate the relationship between attributes. Firstly, we determine their correlation with the Pearson correlation coefficient

and secondly assess their pairwise scatter plots. This enables us to determine a suitable distance measure and to identify potential cluster tendencies (see Section 4.3.2 to 4.3.4).

1. evaluating the relationship between attributes: creating a scatter plot and calculating the binary linear correlation
    a. determining the distance measure: If a high linear correlation exists eliminate one of the attributes and apply Minkowski Distance (see Table 4.3) or if both have to be within the cluster apply the Mahalanobis distance (see Formula 12, Table 4.3) for cluster algorithm. (see Section 4.3.3)

Attribute weighting

As we do not have any information on the number of cluster nor on which areas are likely to be within a cluster, we can only apply the attribute variance-to-range ratio weighting (see Figure 4.6). This weighting can enhance the cluster structure. However, as we do not know with certainty if the variance of an attributes goes along with the intrinsic groups we also use the standardisation based on range (Milligan & Cooper, 1988), which ensures that each attribute is equally weighted when using the Minkowski distance.

2. transforming the data set given two approaches:
    a. weighting based on variance-to- range ratio (Steinley & Brusco, 2008b) (see Formula 6, Section 4.3.2)
    b. standardising based on range (Milligan & Cooper, 1988) (see Formula 14, Section 4.3.3)

Clustering technique

Given that we do not know the number of clusters in advance we apply clustering techniques which automatically determine the number of clusters. Those techniques are easy and quickly to apply. Therefore, we advise to use all possible techniques and subsequently select the one with the highest validity (see evaluation step below). However, first we have to assess if attributes are normally distributed with the normality test at SPSS. If yes, the assumptions of all automatically clustering techniques (simple EM, x-means, TwoStep clustering) are satisfied, and all can be applied. If not, we apply only the TwoStep clustering and x-means. In contrast to simple EM clustering they do not only rely on likelihood measurements, which is in those algorithms based on a normal distribution, but also on principles of non-model based techniques which makes them more robust (IBM, 2016). By using TwoStep as well as k-means we apply different techniques (hierarchical clustering, k-means clustering), which both have their advantages and disadvantages given a certain cluster structure (see Section 4.3.4). We will evaluate if the outcome of the clusters are suitable for our aim by assessing the number of clusters and their proportions.

Given the risk that these techniques can still result in unreasonable clusters as their assumptions on normality might not fully be satisfied, we also conduct manual clustering. Given that we are not sure on the cluster shape, but examined the data on outliers and noise, we apply two different techniques. For the irregular shape we use single-link clustering and for globular shape k-means (see Figure 4.10). In order to determine the number of clusters we apply the “knee approach” with SSE for k-means clustering and the agglomeration coefficient for single-link hierarchical clustering (see 4.3.5).

Finally, we will conduct a clustering based on a practical approach, without any cluster algorithm or optimisation criterion as within cluster analysis there is always the risk that cluster algorithm might be misleading (see Section 4.3).

3. identifying the number of clusters as well as the cluster assignment automatically; but before, testing if the attributes are normal distributed
    a. if all are normal distributed, applying EM, TwoStep clustering and x-means
    b. if not, applying only TwoStep-clustering and x-means

4. applying single-link hierarchical clustering and k-means; determining the number of clusters by the “knee” approach
5. applying a practical approach by considering the pairwise relations between attributes and the characteristics of the mail delivery process

Evaluation

In order to select the best outcome of those three approaches, we will first assess the internal validity, and if it is sufficient, we will conduct external validation.

For internal validation, we will compare the different clusterings based on their compactness and separation. Therefore, we will apply the Silhouette coefficient, which measures the average distance within a cluster compared to between clusters and indicates the overall compactness and separation (Berkhin, 2006; Tan et al., 2005b) (see Section 4.3.5). For the benchmarking model compactness is more important than separation. Due to the large number of objects (1265 PC5 areas) calculating the average distance between all objects of one cluster (see Formula 16 Section 4.3.5) is too computationally complex. Therefore, we will evaluate the compactness by measuring the distance of objects towards the cluster means instead (sum of squared error (SSE)) keeping in mind that the k-means algorithm probably results in a lower SSE than hierarchical clustering as SSE is incorporated in the objective function (see Section 4.3.4). In order to assess if the new applied clustering techniques lead to an improvement we also calculate the SSE and SC for the original clustering (see Appendix XII).

For the external validation we use quantitative methods including the Rand statistic measure, Jaccard Index and Fowlkes-Mallow as well as a qualitative method by consulting an expert team (Section 4.3.5)

6. evaluating and selecting the clustering output based on internal validation criteria Silhouette coefficient (see Formula 21, Section 4.3.5) and SSE (Formula 15, Section 4.3.4) and compare it to the original cluster division.
7. if internal validation is satisfying, validating externally by an expert opinion and the Rand statistic measure, Jaccard Index and Fowlkes-Mallow Index (Formula 24, 25 & 26 Section 4.3.4)

6.2. Results and Discussion

In the following we present for each step of cluster analysis our results.

6.2.1. Attribute Relation and Distribution

The potential cluster attributes are minor-route per APN, main-route, APN/km², interdrop and APN per meter of main-route (APN/m). We evaluate the attributes given three scenarios: without the hit-chance (Scenario I), peak (Scenario P) and off-peak (Scenario O). Before using the data set, we have evaluated the extreme values (given the box plot, see Appendix X and single link clustering) and have removed those which are due to wrong data and could not be corrected, reducing the sample size from 1265 to 1165 PC areas. PC5 areas excluded are low as well as high populated areas (APN/km²) and contain short as well as long interdrops. More areas are excluded with a lower APN/km² than with an extreme high one, which goes along with the general distribution of our sample BG Utrecht, were the majority of PC5 areas have an APN/km² below 3000, which we will also see later on in this analysis (for instance Table 6.9).

To assess the linear correlation we compare the Pearson correlation coefficient and the statistical significance for all possible pairs (De Veaux et al., 2012) using SPSS. Given the SPSS output (see Table_Apx X-) we can say with a statistical significance (all p-values are lower than 0.05) that all pairs have a correlation. There are four exceptions, the correlation between minor-route and interdrop during peak (given all scenarios) as well as interdrop during off-peak days and main-route, which have no statistical significance ($p>0.05$) (see Table 6.1).

Overall, the correlation between the potential cluster attributes is low (Pearson correlation coefficient < 0.4), except for APN/km² and APN/m, which indicates with a correlation coefficient between 0.67-0.82 a high positive correlation independent of the scenario (see Table 6.1). Therefore, we consider regarding further analysis if one can replace the other and thus use three clustering attributes. For the remaining attributes, we can use Minkowski distance as no strong linear correlation exists. In particular we apply Euclidean distance as it is simple and the most used one within clustering.

TABLE 6.1: NONSIGNIFICANT OR HIGH CORRELATIONS BETWEEN ATTRIBUTES GIVEN SCENARIO I,P AND O

Correlations

		Minor-route	Main-route	i_APN/km²	p_APN/km²	o_APN/km²
i_Interdrop	Pearson Correlation	0.019	.173**	-.184**	-.237**	-.251**
	Sig. (2-tailed)	0.523	0.000	0.000	0.000	0.000
	N	1165	1165	1165	1165	1165
i_APN/m	Pearson Correlation	-.109**	-.193**	.818**	.727**	.688**
	Sig. (2-tailed)	0.000	0.000	0.000	0.000	0.000
	N	1165	1165	1165	1165	1165
p_Interdrop	Pearson Correlation	-0.012	.062*	-.088**	-.116**	-.123**
	Sig. (2-tailed)	0.670	0.035	0.003	0.000	0.000
	N	1165	1165	1165	1165	1165
p_APN/m	Pearson Correlation	-.126**	-.245**	.788**	.728**	.697**
	Sig. (2-tailed)	0.000	0.000	0.000	0.000	0.000
	N	1165	1165	1165	1165	1165
o_interdrop	Pearson Correlation	-0.014	0.046	-.071*	-.095**	-.101**
	Sig. (2-tailed)	0.627	0.114	0.015	0.001	0.001
	N	1165	1165	1165	1165	1165
o_APN/m	Pearson Correlation	-.128**	-.256**	.744**	.694**	.668**
	Sig. (2-tailed)	0.000	0.000	0.000	0.000	0.000
	N	1165	1165	1165	1165	1165

**. Correlation is significant at the 0.01 level (2-tailed).
*. Correlation is significant at the 0.05 level (2-tailed).
relevant for Scenario i (infrastructure)
relevant for Scenario p (peak day)
relevant for Scenario o (off-peak day)

Given the scatterplot of each pair of attributes (see Figure 6.2), we can make the following observations; considering the interdrop and APN/km², we identify two tendencies: For high populated areas the interdrop is generally low: everything higher than 1500APN/km² tends to have less than 15m interdrop. Furthermore, we can see that extreme low populated areas (less than 1000 APN/km²) have mostly more than 4m interdrop, however it can still vary between 4 to 450m. Those findings can be supported by practical reasoning: In highly populated places, like city centres, houses are standing directly next to each other and are often built in the height rather than in the width, resulting in a low interdrop at highly dense areas. In contrast, lower populated areas, houses tend to be wider and more distant to each other. However, this distance often varies (for instance farming houses or two household houses) and therefore the interdrop is highly varying as well.

We can see that there are a few extreme interdrop values (250-1000m) given an APN/km² close to zero (see Figure 6.2). Analysing those PC5 areas (see Table_Apx X-) shows that except the low APN/km², hit-chance (range of 0.8%) and interdrop (range of 750m) are highly varying. Thus within that group we do not expect a cluster tendency. This can be supported by our observation during delivery tours. Furthermore, an expert team of postmen, responsible for delivering highly rural areas, and the senior manager of logistic strategy of delivery confirm those findings. Finally, given the information of the costing and economics department, around 3% of all delivery points are such extreme areas. Thus, it is not only a highly varying but also relative small group and therefore we address them as outliers and discard them for the further cluster analysis.

Considering Scenario P (peak) and O (off-peak) (see Section 6.1), we can see the same pattern between interdrop and APN/km², however the scale is varying. In Scenario P the interdrop is slightly

larger and the APN/km² tends to be smaller than in Scenario I (infrastructure). This difference can be explained with the impact of the hit-chance, which is on average 80% during peak days. Given the formula of APN/km² and interdrop (see Table 5.7), we can see that if the hit-chance decreases, the interdrop increases while APN/km² decreases. With an average hit-chance of 41% during off-peak, we can see a shift of the interdrop upwards and an extreme decrease of APN/km² in Scenario O.

Considering the scatterplot between minor-route and APN/km², we can see that highly populated areas (Scenario I APN/km² > 7000) tend to have less than 4m of minor-route. However, looking at less populated areas, there is no clear tendency between APN/km² and minor-route. In Scenario P and O the expected minor-route stays the same, however the scale of APN/km² is changing; during peak day it gets more spread while on off-peak days the points get more concentrated.
Looking at minor-route and interdrop, we can see that there is no clear cluster tendency, especially at an interdrop less than 20, which is the majority of PC5 areas (89%), the minor-route is highly varying. This can be supported by practical reasoning: given a low interdrop, we can think about flats with stairs, where the minor-route would be quite high in contrast to flats with a mail-box front which would result in zero minor-route. Furthermore, there is no limitation of PostNL on the mail-box location, except that it should be outside and not more than 15m away from the street. Thus while interdrop and APN/km² show a clear tendency, combining it with the minor-route might lead to less distinctive clusters as it does not contribute to the intrinsic grouping of the data, which leads us to following hypothesis:

H3: The variance of the minor-route is independent of the intrinsic grouping of the data and thus does not contribute to a more distinctive clustering.

Therefore, we apply clustering on two different subsets of attributes; one with and the other without the minor-route to which we refer to as Subset 1 and Subset 2 from now on.

Considering the main-route per PC5 there is no clear tendency in the context of the other attributes (see Figure_Apx I-1), which indicates that it should be incorporated in a different way. Therefore, we assessed the number of APN per meter of main-route (APN/m), which as shown in Figure_Apx X-19 results in highly similar patterns like APN/km². Above, we identified a high positive correlation 0.67-0.82 between APN/m and APN/km² which can explain the similar patterns. As using both would not contribute to new information, we exclude one of them and can thus also apply the Euclidean distance. Using APN/m will still provide the same cluster tendencies and gives the advantage that the main-route is directly considered as well. Therefore, we select APN/m instead of APN/km² as cluster attribute.

Finally, comparing the different scenarios, it becomes clear that incorporating the volume, especially on peak days, does not damage the intrinsic cluster pattern of the infrastructure. Therefore, we can reject H1 and assume that clustering by incorporating the hit-chance does not significantly influence the outcome.

Considering that the Pearson correlation coefficient between peak and off-peak day for each attribute is higher than 0.99 (see Table_Apx I-4), we conducted a linear regression to determine their relationship (see Appendix VII – Test Linear Regression Peak and Off-Peak). The result shows that between peak and off-peak the interdrop, APN/m and APN/km² have a significant high linear correlation. However, considering the different dispersions of the data points in the scatterplots of off-peak and peak days, we will apply cluster algorithms for both scenarios separately. Subsequently, we can compare if the PC5 areas are in the same clustering during peak and off-peak days.

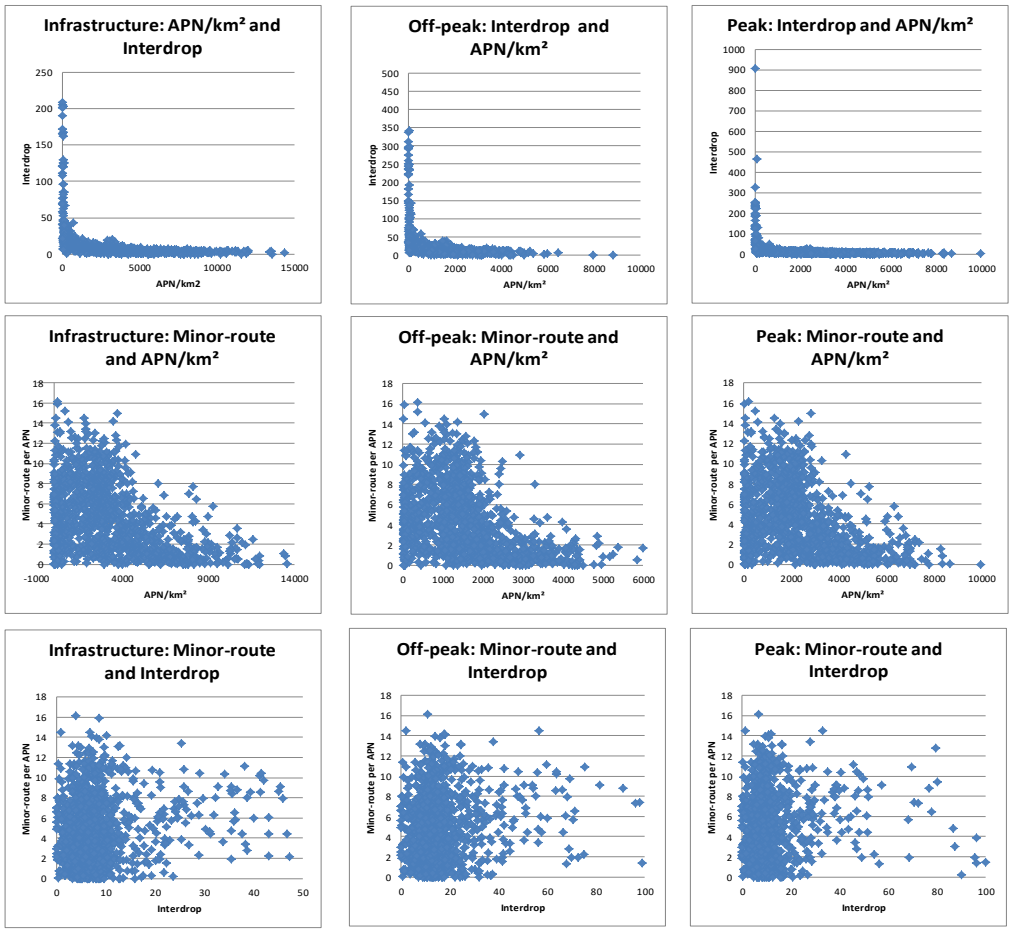


FIGURE 6.2: PAIRWISE RELATION BETWEEN ATTRIBUTES

To summarise, we proceed with APN/m, interdrop as well as minor-route as cluster attributes. As the minor-route does not show a clear cluster tendency, we will apply and evaluate clustering of two subsets. Subset 1 includes APN/m, interdrop and minor-route, while Subset 2 includes only APN/m and interdrop. Finally, we still test the two scenarios, one is clustering the attributes given a peak day (p), and the other is given an off-peak day (o).

6.2.2. Attribute Weighting

Combing two possible attribute subsets (Subset 1: interdrop, APN/m and minor-route; Subset 2: Interdrop and APN/m) and the two scenarios (p = peak, o = off-peak) results in four different cases. The next step is to weight the attributes given the different cases. Therefore, we use the variance-to-range ratio (Steinley & Brusco, 2008b) which is based on the cluster index (CI) of each attribute (see Formula 1, Section 4.3.2) as well as the relative clustering index (RCI) given the subset (see Formula, 3, Section 4.3.2). Table 6.2 and
Table 6.3 present the CI and RCI for each attribute as well as the weights that they gain so that the RCI holds within the transformed space (see Formula 6, Section 4.3.2), where all values are standardised by the z-score (see Formula 4, Section 4.3.2). An RCI of one indicates that the attribute is the least clusterable of all attributes, whereas the RCI of the remaining attributes indicate how many times they are more clusterable than the least one.

Given attribute Subset 1 (S1), the minor-route has in both cases (peak (1p) and off-peak (1o)) the best variance to range ratio. During peak days the minor-route gains 4 times more weight than APN/m and interdrop (see RCI). During off-peak days the CI of minor-route stays constant as it is volume independent, while the CI of interdrop and APN/m changes slightly. The weight is with 4.7 still extremely higher than for APN/m and interdrop. However, as shown in the pairwise analysis the

variance of the minor-route does not seem to support any intrinsic cluster structure (see Figure 6.2), therefore this weighting approach is unsuitable for our clustering and thus will not be applied for Subset 1.

Considering Subset 2, the relative clustering index between APN/m and Interdrop stays the same as in Subset 1. In both cases (1p, 1o) there is nearly no difference in the weight proportion; during peak-days it is nearly the same with both a RCI of 1 and during off-peak APN/m gains 0.1 more weight than interdrop. Given the only small weighting difference between peak and off-peak (0.1) in Subset 1, we do not expect a high change in clusters during off-peak and peak due to this weighting approach.

TABLE 6.2: SUBSET 1: THREE ATTRIBUTES (CASE 1P = GIVEN PEAK DAYS, 1O= GIVEN OFF-PEAK DAYS)

S1	Peak			Off-Peak		
	CI	RCI	Weighting in transformed space	CI	RCI	Weighting in transformed space
Interdrop	0.13	1.0	0.0028	0.12	1.0	0.002
APN	0.12	1.0	0.0026	0.13	1.1	0.003
Minor-route	0.54	4.4	0.0246	0.54	4.7	0.025

TABLE 6.3: SUBSET 2: TWO ATTRIBUTES (CASE 2P = GIVEN PEAK DAYS, 2O= GIVEN OFF-PEAK DAYS)

S2	Peak			Off-Peak		
	CI	RCI	Weighting in transformed space	CI	RCI	Weighting in transformed space
Interdrop	0.13	1.0	0.0014	0.12	1.0	0.0012
APN	0.12	1.0	0.0013	0.13	1.1	0.0014

Considering the quite extreme weighting of Subset 1 and the fact that the variance of an attribute not always contributes to the intrinsic cluster structure, we decide to conduct the clustering with equal weights. As we use Euclidean distance, we apply the standardisation by range, which ensures that each attribute has a range of 1 (see 4.3.2). As the weight proportion in Subset 2 is nearly equal, we do not expect a different outcome than by standardising it. Therefore, we do not consider Case 2a and 2b any further, leading to the final cases presented in Table 6.4.

TABLE 6.4: CASES ON WHICH THE CLUSTERING TECHNIQUES ARE APPLIED

Case	Description
1p	Peak-day, given standardized attributes Interdrop, APN/m, minor-route
1o	Off-Peak-day, given standardized attributes Interdrop, APN/m, minor-route
2p	Peak-day, given standardized (=equal weighted) attributes Interdrop, APN/m
2o	Off-Peak-day, given standardized (=equal weighted) attributes Interdrop, APN/m

6.2.3. Clustering Technique

To determine clusters, we apply a theoretical as well as a practical approach. The former defines clusters based on the cluster algorithm, whereas the latter is based on manual clustering given the attribute relationships. As each clustering algorithm is based on different optimisation criteria (automatically clustering has the BIC/AIC coefficient and k-means the SSE), we expect different cluster outcomes. For the practical approach we do not use an algorithm with an optimisation criterion on purpose as those not always go along with the natural patterns (see Section 4.3.1).

We will first discuss the results of automatic clustering (TwoStep clustering, x-means), secondly of clustering techniques with the number of clusters based on the "knee approach" (single-link

hierarchical, k-means) and finally of the practical approach. In Section 6.2.4 we will compare and select the best cluster outcome of those different approaches.

Automatically Clustering

Given the output of the normality test of SPSS (see Appendix X - Normality Assessment), we can assume with statistical significance that none of the attributes is normally distributed. Therefore, algorithms sorely based on model-based approaches assuming a Gaussian distribution should not be applied, which includes simple EM. However, algorithms like x-means and TwoStep clustering of SPSS combining model-based approaches with principles of non-model based techniques (k-means, hierarchical clustering) are more robust and thus can still be applied.

In Table 6.5 we present the outcome of the cluster algorithm and show the number of clusters, the cluster proportion and the Silhouette coefficient (SC) for each scenario. As mentioned above, the SC can be between -1 and 1. The closer the SC is to 1, the better the clustering.

We have differentiated between the BIC and AIC criteria within TwoStep clustering to ensure reliability. Looking at their outcome (see Table 6.5), all clusters have a SC higher than 0.5 indicating a good cluster outcome considering the compactness and separation. However, we can see that the TwoStep clustering forms a cluster with at least 97% of all PC5 areas in every scenario, independent if we use the BIC or AIC coefficient, and thus should not be used as an orientation for the number of clusters.

TABLE 6.5: OUTCOME OF THE 2STEP AND X-MEANS CLUSTER ALGORITHM

	TwoStep (Euclidean distance)						x-means (max 100 iterations, Euclidean distance)		
	BIC criterion			AIC criterion			BIC criterion		
Case	Number of clusters	cluster proportion (%per cluster)	SC	Number of clusters	cluster proportion (%per cluster)	SC	Number of clusters	cluster proportion (%per cluster)	SC
1 p	3	97, 2, 1	0.6	3	97, 2, 1	0.6	2	59, 41	0.61
o	3	96, 2, 2	0.6	3	96, 2, 2	0.6	2	59, 41	0.61
2 p	3	97, 2, 1	0.9	3	97, 2, 1	0.9	4	68, 28, 3, 2	0.53
o	3	96, 2, 2	0.8	3	96, 2, 2	0.8	4	66, 29, 2, 2	0.54

The x-means algorithm proposes more equally sized clusters for Subset 1 with a slightly lower Silhouette Coefficient of around 0.52, however it is still reasonable. Looking at Subset 2, the number of clusters increases to 4 with two smaller clusters (2% and 3% of the PC5 areas incorporated).

The different cluster numbers and proportions of Subset 1 and 2 in x-means clustering are due to the attribute minor-route. Whereas it is not considered in Subset 2, minor-route is a determining attribute in Subset 1. There is a clear separation between the clusters within the minor-route and APN/m plot as well as the minor-route and interdrop plot (see Figure 6.3), but an intermingling within the interdrop and APN/m plot (see Figure 6.4). Excluding the minor-route leads to clearer division within the interdrop and APN/m plot; in Subset 2 the extreme cases of interdrop (for 2p > 73m, for 2o > 101m) and APN/m (for 2p > 0.16 APN/m, for 2o > 0.10APN/m) gain a separate clusters, increasing the number of clusters from 2 to 4 (see Figure 6.4).

Finally, we can see a high similarity between the cluster number and proportion between peak and off-peak days and by that disconfirm H2. Conducting a paired sample t-test based on the cluster membership (see Figure_Apx XI-1) shows that there is no significant difference in the membership regarding Subset 1, all PC5 areas are within the same cluster during peak and off-peak days. However, in Subset 2 3% of PC5 areas are assigned to different clusters in peak and off-peak days. Given the paired sample t-test, we can confirm that there is a significant difference, even though it is a small difference with an absolute mean of 0.043. This small difference can be explained by the high linear relationship linear relationship between peak and off-peak days of the attribute interdrop and

APN/m, meaning that for all PC5 areas the values increase or decrease uniformly (see Appendix VII - Test Linear Regression Peak and Off-Peak). Given that in Subset 1 the minor-route is a determining factor for the cluster assignment and constant during peak and off-peak, it can explain that there is no significant difference in Subset 1 at all.

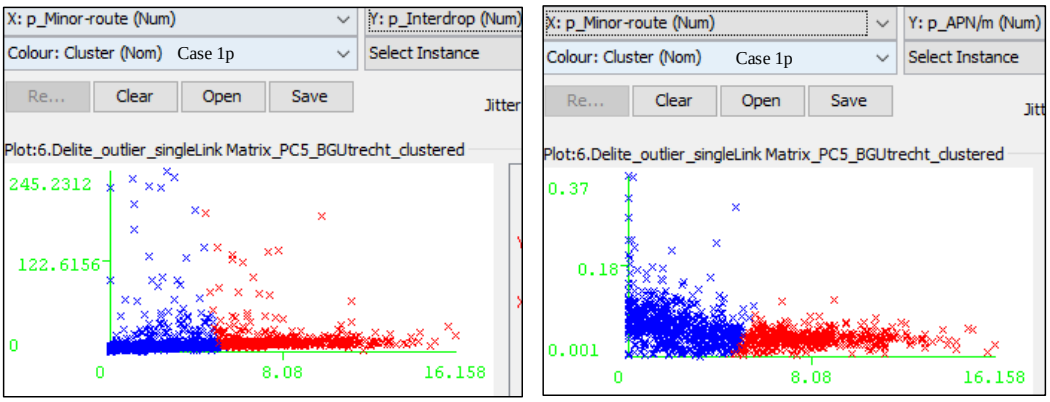


FIGURE 6.3: SCATTERPLOT IN WEKA (MINOR-ROUTE VS INTERDROP (IN THE LEFT), MINOR-ROUTE VS APN/M (IN THE RIGHT)) FOR CASE 1P

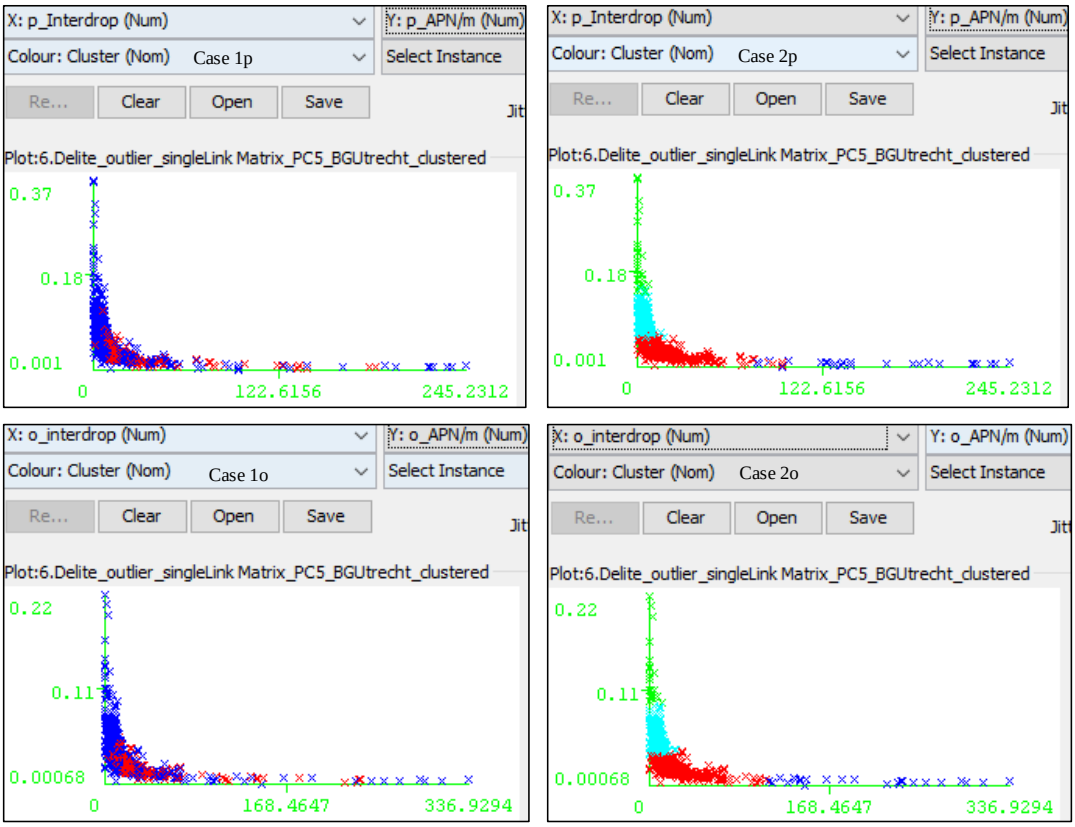


FIGURE 6.4: SCATTERPLOT IN WEKA: INTERDROP VS APN/M FOR CASE 1P (TOP LEFT),1O (BOTTOM LEFT), 2P (TOP RIGHT) AND 2O (BOTTOM LEFT) WITH THE COLOUR INDICATING THE CLUSTER MEMBERSHIP

Finally, the difference in cluster proportions between TwoStep and x-means clustering can be explained by their underlining clustering technique. While TwoStep clustering uses hierarchical clustering, which incrementally creates clusters by grouping the closest objects together (based on a centroid linkage), x-means clustering is based on k-means approach, which directly starts with a certain number of clusters and iteratively adds objects and adapts centres based on the SSE.
Analysing the hierarchical clustering, we can see no high jumps in the agglomeration coefficient within the agglomeration table, except in the last 2 to 4 merges, indicating that in the previous steps

only high similar clusters are merged. This is also visualised by the dendrogram of each case. Even before the last 4 merges the cluster proportions are highly unequal with one extreme big cluster. (see Appendix XI – Dendrogram of Hierarchical Clustering)
Thus we conclude that TwoStep clustering is with its cluster proportion unsuitable for our aim. Therefore we will only proceed with the outcome of x-means as potential clustering.

Clustering Techniques with K as Input

In this section we will determine the number of clusters with the "knee" approach. Therefore, we will apply single-link as well as k-means clustering with the Euclidean distance measure with k-varying from 1-12 for each of the four cases.

In all cases single-link clustering places at least 96% of all PC5 areas in the same clusters (see Appendix XI - Single-Link Hierarchical Clustering), indicating that there are no distinctive groups in the sense that there are no significant separations between objects. In order to ensure that this major cluster does not come due to a "bridge of noise" (see Section 4.3.4) we have visualised the clustering using WEKA and have not identified any bridges, but instead can support the finding as looking at the scatter plot interdrop - APN/m there is one continuous sphere of objects (Figure 6.2). Same accounts for the scatterplot minor-route - interdrop and minor-route - APN. Overall, the findings of single-link clustering are consistent with the findings of TwoStep clustering, which also relies on hierarchical clustering. Thus, we can conclude that creating highly distinctive cluster is not possible; however, we can still try to ensure high similarity within the clusters by applying the k-means algorithm, which is sorely based on the SSE per cluster and thus the internal cluster homogeneity.

To ensure comparability between the results of k-mean algorithm with different number of clusters, we use the same random seed number (seed 10 in WEKA) and the maximum number off iteration (500). None of those k-means runs has required more than 100 iterations, thus a maximum of 500 does not limit the optimisation procedure of the k-means algorithm. We have increased k until there is no significant change of the average SSE within a cluster.

In both subsets the graphs during peak and off-peak do not show a difference, which supports the findings of the x-means clustering that the cluster structure during peak and off-peak is similar (see Figure 6.4).

For Subset 1, the "knee" would be at k equal to 6 (see Figure 6.5). Same accounts for Subset 2 (see Figure 6.5). Comparing the scale of the SSE between the scenarios, we can see that Subset 2 shows much compacter clusters (SSE between 0-20) than Subset 1 (SSE between 0-40) given the same k. This indicates that adding the minor-route as attribute leads to less compact clusters. Later on, in the evaluation, we apply more validation indices to make a final conclusion.

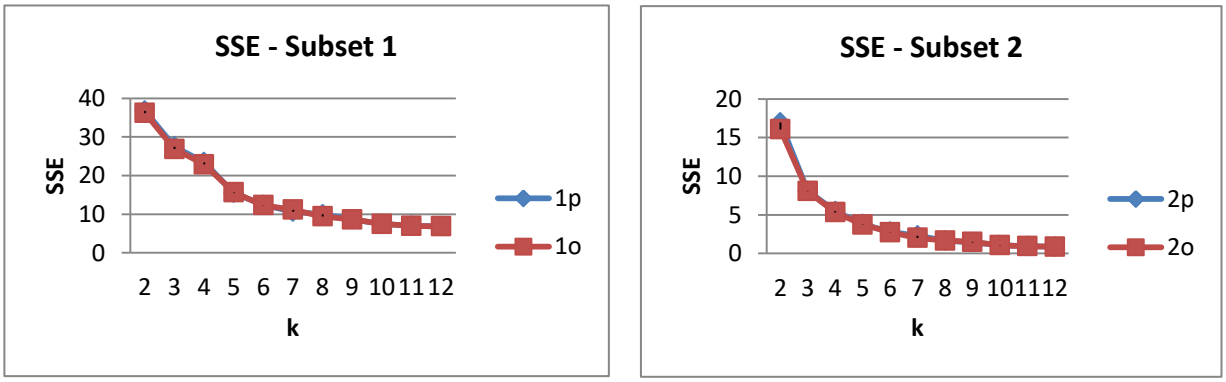


FIGURE 6.5: SSE WITHIN A CLUSTER PER K-CLUSTER GIVEN SUBSET 1 AND 2

The next step is to apply the k-means algorithm with k equal to 6. To minimise the risk of local minima we have run the k-means algorithm with different randomly placed initialisation points and

have compared their SSE. For all four scenarios the SSE has not changed significantly; within Subset 1 there has been a maximum difference of 0.14 given 10 different random initialisations and a minimum SSE of 12.02 (Case 1p) and 12.22 (Case 1o). In Subset 2 the difference between the SSE is 0.01 and has a minimum SSE of 2.8 (Case 2p) and 2.72 (Case 2o). The small difference in SSE within the 10 random initialisations indicates that there are probably no extreme local optima. Secondly, the high difference between Scenario 1 and 2 of 10 SSE indicates that the clusters are much more compact when omitting the minor-route as cluster attribute.

The final clusters for each case are defined in Table 6.6, which shows the number and percentage of objects per cluster and their cluster centroids. The r in front of the attribute name indicates that the attributes are standardised by their range, while the o or p indicates the case (peak or off-peak). For instance in Case 1p the cluster 1 incorporates 28 objects, which are 2% of all objects.

TABLE 6.6: CLUSTER INFORMATION FOR EACH CASE (NUMBER AND PERCENTAGE OF OBJECTS WITHIN EACH CLUSTER AND THEIR CLUSTER CENTROIDS) BASED THE K-MEANS ALGORITHM (WITH K=6)

Case 1p

	Cluster number						
Attribute	Full	1	2	3	4	5	6
	1140 (100%)	28 (2%)	26 (2%)	282 (25%)	125 (11%)	304 (27%)	375 (33%)
r_p_Interdrop	0.0607	0.6181	0.0074	0.0561	0.0568	0.0609	0.0273
r_p_APN/m	0.1438	0.0176	0.5724	0.1048	0.1012	0.1084	0.1956
r_p_minor-route	0.2885	0.2309	0.0671	0.456	0.6918	0.2507	0.0785

Case 1o

	Cluster number						
Attribute	Full	1	2	3	4	5	6
	1140 (100%)	25 (2%)	30 (3%)	282 (25%)	125 (11%)	300 (26%)	378 (33%)
r_o_Interdrop	0.0624	0.622	0.0104	0.0596	0.0596	0.0635	0.0317
r_o_APN/m	0.1553	0.0217	0.569	0.1158	0.1135	0.1182	0.2039
r_o_minor-route	0.2885	0.2446	0.0782	0.456	0.6918	0.2525	0.0784

Case 2p

	Cluster number						
Attribute	Full	1	2	3	4	5	6
	1140 (100%)	14 (1%)	129 (11%)	354 (31%)	563 (49%)	60(5%)	20 (2%)
r_p_Interdrop	0.0607	0.0029	0.0177	0.0273	0.0527	0.2216	0.712
r_p_APN/m	0.1438	0.6909	0.2999	0.1727	0.0923	0.0348	0.0162

Case 2o

	Cluster number						
Attribute	Full	1	2	3	4	5	6
	1140 (100%)	16 (1%)	128 (11%)	345 (30%)	76 (51%)	59 (5%)	16 (1%)
r_o_Interdrop	0.0624	0.004	0.0206	0.0309	0.0561	0.2301	0.7458
r_o_APN/m	0.1553	0.6846	0.3138	0.1861	0.1022	0.0428	0.0197

Similar to automatic clustering we can see nearly the same cluster proportions between off-peak and peak independent of the attribute subset that we use. Given the significant linear relationship between off-peak and peak attributes (see Appendix VII – Test Linear Regression Peak and Off-Peak), we have also developed a linear model for the standardised interdrop and APN/m (see Appendix XI – Test Linear Regression Peak and Off-Peak for by range standardised attributes). Appling this model to the cluster centroids of Case 2p results in nearly the same cluster centroids as in Case 2o (see Table 6.7). Comparing the minor-route values of the centroids between Case 1p and 1o, we can see that there are also only small differences (maximum difference of 0.013 at cluster 1). Thus, also for this clustering technique we conclude that there is no significant difference between peak and off-peak and reject H2.

TABLE 6.7: CLUSTER CENTROIDS CALCULATED BASED ON THE LINEAR MODEL, GIVEN THE CENTROIDS OF THE PEAK CASES

Case 1o		Cluster number						
	Attribute	Full	1	2	3	4	5	6
linear model	r_o_Interdrop	0.05887396	0.641295	0.003181	0.054067	0.054799	0.059083	0.023975
	r_o_APN/m	0.132659816	0.010614	0.547152	0.094944	0.091462	0.098425	0.182755

Case 2o		Cluster number						
	Attribute	Full	1	2	3	4	5	6
linear model	r_o_Interdrop	0.05887396	-0.00152	0.013944	0.023975	0.050515	0.226996	0.73941
	r_o_APN/m	0.132659816	0.661752	0.283622	0.160609	0.082855	0.027248	0.00926

Practical Approach

Cluster algorithms always find clusters based on a certain criterion, but if intrinsic clusters of the data set does not fit the assumptions of the algorithm (Tan et al., 2005b; Xu & Wunsch, 2005), the clusters outcome might not be representative. Therefore, in this practical approach we create clusters without an algorithm and optimisation criterion but based on the relationship between attributes and the current strategy of PostNL.

The first main division is based on the means of transportation as they highly influence the velocity of the delivery. Currently, PostNL selects the means of transportation based on their own developed

model, which shows the most cost efficient transportation means per interdrop. In the previous section, we therefore have considered the interdrop as an attribute, but have not directly clustered based on the means of transportation. This is due to three reasons. Firstly, we assume that if clusters have the same interdrop, they should have the same means of transportation. Secondly, the current optimal means of transportation per interdrop can change, if the costs for the means of transportation decrease. Hence, clustering based on interdrop provides a more long term applicable solution. Thirdly, the interdrop boundaries for the means of transportation might not be the boundaries of the intrinsic clusters and thus might hamper that cluster algorithms find clusters.

The selection of means of transportation is based on the interdrop during peak days because the majority of delivery days are peak days (3 of 5 days). Furthermore, per delivery tour they assign only one means of transportation, thus they have to ensure that the means of transportation pay off during peak as well as off-peak days. With an increase in the interdrop, the costs of the means of transportation increase: for instance the car has to be paid by PostNL which is used for a large interdrop, while for bike deliveries postmen use their own bike leading to no costs for PostNL. Thus, selecting the means of transportation during peak days, which has a smaller average interdrop than off-peak days, ensures that it will always pay off.

Therefore, we split the PC5 areas in groups based on the turning points for means of transportation given the interdrop suggested by the model of PostNL. Due to confidentiality, we only name a range for the turning points instead the exact point in this paper. This results in following groupings: an interdrop less than 20-40m (foot/bike), between 20-70m (bike, e-bike), between 50-120m (scooter) and higher than 100-120m (car) interdrop during the peak days (see Table 5.5). Bike and e-bike delivery are grouped together as the current number of e-bikes is small with 200 bikes in the whole Netherlands. However if the number increases, one should reconsider the grouping.

For the second division we consider the relationship between interdrop and APN/km² (see Figure 6.6), which is often used as an indicator for the population density or degree of urbanisation and thus easy to interpret for users of the benchmarking model. As we have shown in Section 6.2.1, it has a high linear relationship with APN/m and thus using APN/km² instead of APN/m does not lead to an information loss.

Within the groups based on the means of transportation we can identify patterns. For car deliveries (interdrop above 100-120m) the majority of PC5 areas has less than 50 APN/km² except a few outliers which do not exceed 100 APN/km². For PC5 areas with scooter deliveries, the APN/km² is mostly below 100 APN/km² with a few exceptions that are not above 150 APN/km². PC5 areas with bike and e-bike deliveries, the APN/km² is mostly below 400 APN/km² with some exceptions that are not above 700 APN/km². However, for foot/bike deliveries (less than 20-40m interdrop), which is around 91% of all deliveries (see Table 6.9), the APN/km² of PC5 areas varies highly (0-15000APN/km²) (see Figure 6.6) and therefore this group requires some further divisions based on the APN/km².

Three natural divisions can be found: firstly, between 0 and 1000 APN/km² as the interdrop is mostly above 5m and above 1000APN/km² the interdrop is mainly less than 20m (see Figure 6.5 and Appendix XI - Practical Clustering Approach). Secondly, at an APN/km² of 3000 as the majority of PC5 areas above that have less than 8m minor-route. Finally, at 5000 APN/km² as from there the majority of PC5 have less than 5m minor-route (see Figure 6.5). Overall, this results in a division of four groups for PC5 areas with an interdrop of less than 20-40m (see Table 6.9). Looking at the proportion of the four groups (see Table 6.9) we can see that cluster 5 is highly dominating with 47%. As we do not know with certainty if this proportion only accounts for our sample BG Utrecht, we keep it by the three natural divisions for now. However, after having applied this clustering to the whole Netherlands one should recheck the proportions and if cluster 5 is still domination, we advise to split it into two equally sized clusters.

Overall, with the practical approach we gain seven clusters (see Figure 6.7). The division based on the interdrop is due to confidentially not explicitly shown, instead only a possible range for the divisions.

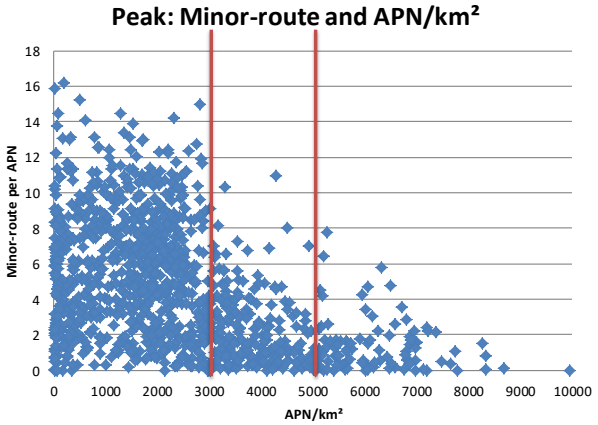


FIGURE 6.6: SCATTERPLOT MINOR-ROUTE AND APN/KM²OF PC5 AREAS

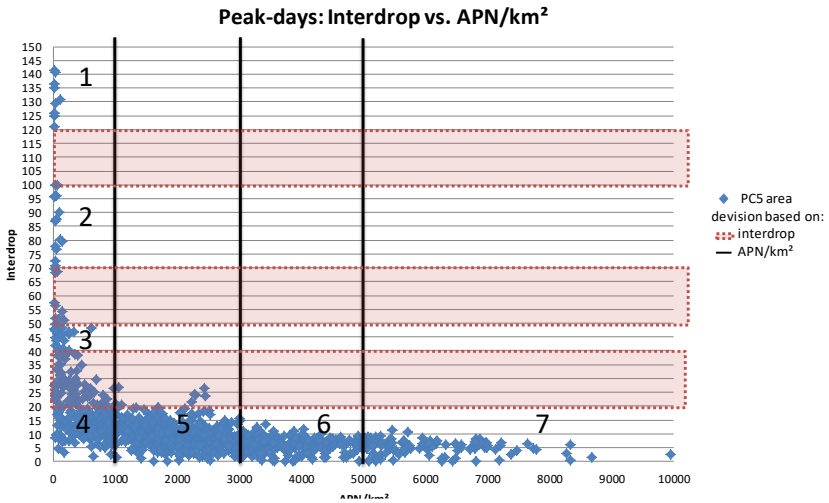


FIGURE 6.7: CLUSTERING OF PC5 AREAS BASED ON A PRACTICAL APPROACH WITH THEIR CLUSTER NUMBER

TABLE 6.8: 1.DIVISION BASED ON INTERDROP

Interdrop (lower limit)	Means of Transportation	Final Cluster	Number of objects	% of objects
100-120	car	1	21	2%
50-70	scooter	2	23	2%
20-40	bike/ e-bike	3	61	5%
0	foot		1036	91%

TABLE 6.9: FURTHER DIVISION OF OBJECTS WITHIN THE "FOOT/BIKE" CLUSTER (INTERDROP < 20-40M)

APN/km²	Final Cluster	Number of objects	% of objects
5000	7	109	10%
3000	6	217	19%
1000	5	538	47%
0	4	172	15%

6.2.4. Evaluation and Conclusion

In order to evaluate and to compare the outcome of the different clustering techniques, we assess the internal validity as well as the external validity. For the former we apply the Silhouette Coefficient (SC) introduced by Rousseeuw (1987) and the sum of squared error (SSE) (Berkhin, 2006;

Tan et al., 2005b). For the latter one we apply with the Rand statistic measure, the Jaccard Index and the Fowlkes-Mallow Index quantitative external validation as well as qualitative validation by consulting an expert-team, which includes the senior controller who is responsible for the benchmarking, the senior process manager of optimisation and the senior manager of logistic strategy with specialisation in mail delivery.

Internal Validity

The quality of clusters is measured based on the SC and SSE. The former one measures the overall separation and coherence of the clustering. The SC can vary between -1 and 1; the closer it is to 1, the better. The latter solely measures the coherence of the clusters. The closer the SSE value is to zero, the smaller the distance between objects within a clusters and thus the higher the similarity within a cluster. In contrast to the SC, there is no upper bound for the SSE.
In order to evaluate if the new approach is an improvement, we also calculate SSE and SC given the original clustering. Within the original clustering delivery areas are defined as cluster objects and divided in five clusters based on the APN/km² (see Table 1.1). As we do not have the data of the other delivery areas (= cluster objects), we cannot calculate the SC (see Formula 21) or the SSE to evaluate the coherence and separation of the overall clustering. However, we can assess the coherency within the BG Utrecht and compare if the coherence significantly improves if we sub-divide BG Utrecht into multiple clusters. By that we can assess if a new clustering is worth the effort. Furthermore, we apply the original cluster division instead of on delivery areas on the PC5 areas in order to assess if the scaling of APN/km² (see Table 1.1) is better than the one of the practical approach (see Table 6.8 and Table 6.9). During the benchmarking sessions, they also differentiate the performance between three categories: car, scooter or remaining deliveries (incl. bike, e-bike and foot). Therefore, we divide the PC5 areas not only based on Table 1.1 but also regarding the three categories and subsequently calculating SC and SSE.

Both criteria, SSE and SC, rely on a distance measure. To ensure a fair comparison, we apply the internal validation criteria to the standardised range of attributes for all techniques (Milligan & Cooper, 1988). Given the same range and a high correlation between APN/m and APN/km² (see Table 6.1), we can compare the practical approach which is based on peak days with Case 2p of the theoretical approaches which considers the interdrop and ANP/km² (attribute Subset 1) during peak days (Scenario P), although the former is based on APN/m and the latter is based on APN/km². For further description of the cases see Table 6.4.
Furthermore, because the results of both cluster algorithms (the automatic and k-means clustering) show no significant difference between peak and off-peak, we can reject our first hypothesis and conclude that there is no need to differentiate between those days for the clustering during the benchmarking. Moreover, this means that we can focus on one scenario for our evaluation and comparison. As the practical approach is only based on peak days, we can compare better by considering the peak case.

Looking at SC (see Table 6.11), we can see that cluster separateness and compactness are for all techniques reasonable with a SC vale around 0.50, except for the original cluster division which performs with a SC of 0.142 poorly. Overall, with 0.52 (Case 1p) and 0.54 (Case 2p) x-means technique performs slightly better than the other techniques. However, considering SSE (see Table 6.10) the compactness of x-means clusters is the lower than k-means or the practical approach; x-means has with 57 compared to 41 (k-means) and 52 (practical) a higher SSE given Case 2p and with 160 compared to 99 (k-means) a higher SSE in Case 1p (see Table 6.12). This means that the distance between objects within a cluster is higher for x-means than any other new presented technique, resulting in less compact clusters. Knowing that the compactness is low, we can conclude that SC, which combines compactness and separation (see Formula 21, Section 4.3.5), is slightly higher than the other due to better cluster separation rather than the compactness. Therefore, x-means is less suitable for benchmarking than the other techniques.

Even though the practical technique is not directly based on distance calculations, it still performs reasonably. With 0.455 the SC is only slightly less (0.04) than the SC of k-means. Considering the SSE of 52 it is better than x-means which has a SSE of 57, but worse than k-means. K-means achieves the lowest SSE, which is 27% lower than the one of the practical approach and thus delivers the most compact clusters. Given that the objective function of k-means is to minimise SSE, we have already expected a lower SSE. However, setting the 27% in relation to the difference to other scores of SSE, this difference is the smallest. Furthermore, knowing that the SSE for the k-means approach would not improve significantly anymore with k equal to 7 (see Figure 6.5), the practical technique with seven clusters has performed reasonable with "only" a difference of 27%.

If we only differentiated between car, scooter and remaining deliveries within BG Utrecht, which is the case of the current benchmarking, we would gain a SSE of 102 (see Appendix XII for an outline of the calculation). Applying x-means (SSE of 57), k-means (SSE 41) or the practical approach (SSE 52) would improve the cluster compactness with around 50%. Therefore, it is worth to sub-divide BG Utrecht further.
Applying the current APN/km² scaling (see Table 6.11) on PC5 areas does still result with 88 in the highest SSE (see Table 6.12). This is mainly because the current division from 0 to 1000 APN/km² is unsuitable for PC5 areas as cluster objects. As discussed in Section 1.3, delivery areas contain areas with highly different household densities (APN/km²). Instead of taking the average of a delivery area, measuring the APN/km² per PC5 area is more precise allowing more extreme values, which results in a significant larger range. As a result 76% off the PC5 areas are within the cluster that allows with a boundary of >1000 APN/km² the highest densities (see Appendix XII). Given that we have PC5 areas with up to 10000 APN/km² (see Figure 6.7), we can explain that a cluster with PC5 areas from 1000 to 10000 APN/km² does not show a high coherence resulting in an overall low compactness of the clustering outcome. For more details see Appendix XII.

Finally, comparing Case 1p with Case 2p (Case 1p incorporating Subset 1: APN/m, interdrop and minor-route, Case 2p incorporating Subset 2: APN/m and interdrop), we can see that Case 1p performs worse than Case 2p given all techniques. While SC decreases only slightly, the SSE score is increasing with 280% (x-means) and 240% (k-means) extremely. For x-means we can argue that the change can be due to the difference in the number of clusters, as in general SSE decreases with the number of clusters. However, in k-means the number of clusters stays constant between Case 1p and Case 2p, and thus the 240% increase is sorely due to the minor-route. This confirms our third hypothesis (see Section 6.1) stating that the minor-route does not contribute towards a high cluster quality and therefore should be excluded.

Overall, for the clustering of the benchmarking only the peak case has to be considered. From an internal validity perspective with the focus on cluster compactness k-means clustering has performed best, closely followed by the practical technique.

TABLE 6.11 THE SILHOUETTE COEFFICIENT (SC) FOR EACH CLUSTERING TECHNIQUE AND FOR THE ORIGINAL CLUSTERING APPLIED TO THE PC5 AREAS

SC	x-means	k-means	practical	original
1p	0.52	0.417	-	-
2p	0.54	0.494	0.455	0.142

TABLE 6.12: SUM OF SQUARED ERROR (SSE) FOR EACH CLUSTERING TECHNIQUE AND FOR THE ORIGINAL CLUSTERING APPLIED TO THE PC5 AREAS

SSE	x-means	k-means	practical	original
1p	160	99	-	-
2p	57	41	52	88

External Validation

The external validation is used to assess if a cluster outcome is realistic and if the patterns found can be confirmed. Therefore, we follow two approaches: One is quantitative and is based on the Rand statistic measure, the Jaccard Index and the Fowlkes-Mallow Index, indicating the degree of match between the cluster outcome and external information. The other one is qualitative and implies to ask an expert team to assess the cluster outcome.

To conduct qualitative external validation (see Section 4.3.5), we require external information on the similarity between PC5 areas. Therefore, we ask team leaders to assess for each possible pair of PC5 areas if there is a high similarity and thus should be in the same cluster. Team leaders have an exact overview of the areas that they have to manage because they work on-site and take care of the customers, depots and postmen within that area (see Section 3.1). However, the team leaders indicate that none of the areas and their delivery time could be compared as each had its own characteristics. Therefore, we cannot retrieve the necessary information to conduct a quantitative validation.

In contrast, process managers, the process manager of optimisation and the senior manager of logistic strategy specialised in delivery are not able to say in detail if a certain pair of PC5 areas was similar, however they are able to assess if the cluster division was reasonable. As our clustering is based on a maximum of three attributes the cluster outcome is easy to understand and to interpret (see Section 4.3.5). Therefore consulting an expert team is the easiest method to gain external validation and a reasonable alternative for the quantitative validation.

The expert team has confirmed Hypothesis 3 (see 6.2.1) as it has also not expected a clear pattern in minor-route and interdrop as well as APN/km², arguing that PostNL did not set any restrictions on the mailbox location except that it has to be within 15m from the street. Thus, the high SSE for attribute Subset 1 is reasonable.

Considering Hypothesis 2, the team agrees that the high similarity between the cluster outcome of peak and off-peak was reasonable. According to it, one can give an adequate estimation for off-peak days given the characteristics of a peak day for a certain PC5 area, indicating a high dependence between them.

However, the experts do not agree with the cluster outcome of k- and x-means as within the same clusters some PC5 areas are highly different due to dissimilar means of transportation in their opinion. The expert team argues that if you wanted to create clusters to compare the delivery time per mail item, the primary match between objects should be the means of transportation. Thus, for the expert team the cluster outcome of the practical clustering technique is the most realistic.
Within cluster analysis there is the possibility of using nominal cluster attributes like means of transportation. In Section 4.3.3, we have defined how to measure the distance between those attributes. However, we have not used means of transportation within the algorithms on purpose as explained in Section 6.2.3. The main reason is that the means of transportation imply a managerial decision. The aim of clustering is to ensure that the attributes are similar, thus next to APN/km² the interdrop should be similar too, which is the main selection criteria for selecting the means of transportation. If a cluster still contains different means of transportation although the areas are similar, one could discuss which PC5 area selected the means of transportation most optimally.
However, after evaluating the cluster tendencies, coherence and separation, we could not identify highly distinctive clusters for our performance measure. Hence, the interdrop within a cluster is not highly similar. Therefore, our strategy mentioned above does not function. However, to improve to some extant the similarity and thus a better comparability of the delivery time, we should at least ensure that the means of transportation is the same.

Overall, considering both, internal as well as the external validation, we conclude that the practical clustering technique delivers the most reasonable clustering. According to internal validation the cluster quality is slightly lower than the one of k-means, however it outperforms k-means at the

external validation. Therefore, we advise PostNL to apply the clusters from the practical clustering technique for the benchmarking model.

6.3. Conclusion

In this chapter we have defined and applied the test framework for cluster analysis of the mail delivery performance.

We first has analysed the potential cluster attributes, interdrop, APN/km^2, minor-route, main-route and APN/m of PC5 areas given our sample the delivery area of Utrecht. We have differentiated between three scenarios, peak- and off-peak days as well as infrastructure, in which the number of points that receive delivery differ. We have come to the conclusion that there is no significant difference between patterns in the infrastructure and the peak-day scenario as with an average hit-chance of 80% during peak days the majority of delivery points receive delivery. Furthermore, none of the attributes shows high linear correlation except APN/m and APN/km². Looking at the pairwise scatterplot, we have come to the conclusion that APN/m and APN/km² are interchangeable and thus it is sufficient to only use one of them for the clustering.

Whereas interdrop and APN/m have shown a clear pattern, no pattern could be identified with the attribute minor-route. Therefore, we have tested two different attribute subsets: Subset 1 with interdrop, APN/m and minor-route and Subset 2 with only interdrop and APN/m. Given the internal validation criteria, we can see that Subset 2 performed worse than Subset 1, confirming that the minor-route does not contribute to a better cluster structure; whereas the Silhouette Coefficient of Subset is slightly lower than in Subset 2 with a difference of 0.02 (x-means) and 0.08 (k-means), the SSE is more than 50% higher than in Subset 2 indicating a low compactness of the clusters.

We have tested three different clustering techniques: automatic clustering, clustering with K as input and practical clustering. For clustering techniques using an algorithm, we have applied cluster attributes standardised by range to ensure equal weight. The weighting technique range by variance has not shown a reasonable outcome as it has given the attribute minor-route four times more weight than the other attributes.

Overall, independent of the clustering technique, there has not been a significant difference between peak- and off-peak days as PC5 areas are during peak as well as off-peak are within the same group. Therefore, PostNL does not have to differentiate between those days within its benchmarking model.

Comparing the outcome of the different techniques with each other we can conclude that k-means performance with a SSE of 41 and a reasonable silhouette coefficient the best on internal validity, closely followed by the practical approach with a SSE of 52 and an also reasonable silhouette coefficient. However, given the external validity based on the judgement of an expert team the practical approach outperformances k-means clustering, as according to them the cluster division of the practical approaches ensures the highest similarity between PC5 areas.

Finally, comparing the original clustering to the new ones, we can make two conclusions. Firstly, with a SSE of 102 for the delivery area Utrecht, we can see that using a delivery area as cluster object would result in low coherent clusters. Clustering on a PC5 area would increase the coherency by around 50%, meaning that within the clusters the similarity is significant higher. Secondly, if we apply the current clustering with the APN/km² scaling on PC5 areas it still shows with 88 a high SSE and thus a low cluster compactness. Therefore, the current clustering is not suitable even if the cluster object is on a lower level (PC5 instead of deliver area).

7. Practical Implications and Suggestions for Implementation

To ensure a successful implementation of the benchmarking model we have to consider three aspects. Firstly, the fit between the final clustering derived in Chapter 6 and the managerial structure of PostNL. Secondly, the lay-out of the benchmarking-model and finally the implementation plan to cover the last steps of the benchmarking process.

Fit between Clustering and Managerial Structure of PostNL

In Chapter 6 we have developed clusters which do not encounter the managerial structure of PostNL. As defined in Section 5.4, the information required to calculate our performance measure "time per mail item" has only been provided on a team and postmen level until now. Thus, before we can implement the new clustering and performance measure in the benchmarking model, we have to assess the fit with the managerial structure and to clarify possible implications on the quality of the benchmarking.

The main idea is to compare the same clusters between the different delivery areas. Therefore, ideally we sum up the required delivery time and the weighed mail volume of all PC5 areas that are in the same delivery area and same cluster. Whereas the mail volume is given per PC5 area, the delivery time is currently only given per team and per postman. PostNL has assigned each PC5 area to a depot and each depot to a team. However, as we can see in Figure 7.1, the majority (84%) of teams manages PC5 areas which are in more than four different clusters given the new clustering. As the delivery times of car and scooter areas are booked separately from (e-) bike and foot deliveries, we can extract Cluster 1 (interdrop >100m) and Cluster 2 (interdrop >50m) from the remaining clusters. However, still teams contain different clusters and thus implementing the clustering given the current information availability on team level is only possible by combing different clusters within a team at the expense of the overall cluster homogeneity.

To assess the impact on cluster homogeneity, we evaluate the proportion of the cluster membership of PC5 areas per team. As we can extract the information about the delivery time of cluster 1 and 2, we only consider Cluster 3 to 7 for the cluster proportion. Looking at Figure 7.3, we can see that 90% of all teams incorporate one dominant cluster, meaning that at least 40% of the PC5 areas of a team are in the same cluster. Hence, we can conclude that PC5 areas within a team show a tendency towards one cluster. However, for 72% of the teams Cluster 5 has the highest proportion of all clusters (see Figure 7.2), meaning that Cluster 5 is most often the dominating cluster within a team. This is as expected as 47% of all PC5 areas of delivery area Utrecht are within Cluster 5. Thus, clustering the teams based on their dominating cluster is not only on the expense of the overall cluster homogeneity, but also has the consequence that Cluster 1, 2 and 7 will be empty.

In order to realise all clusters, we advise to couple PC5 areas to the postman that delivers those areas during his delivery tour. The data to do the coupling is available as the tours (including all delivery addresses) of each postman for each day are known, one only has to extract from it the PC5 areas. The advantage is that the number of PC5 areas covered by a postman in on average 2.2, which is significantly less than the PC5 areas of a team with an average of 30.6 (see Appendix XIII). Therefore, the chances of homogeneity between PC5 areas are higher and thus a better cluster quality can be achieved. The disadvantage is that if the tours of a postman vary, it is a higher workload to link a PC5 area to a postman as we have to determine the link for each day separately. However, the majority of postmen always covers the same tour, thus we expect that taking the delivery time per postmen as an information source for our performance measure is realisable and therefore we advise to combine PC5 areas covered by the same postman into one cluster object.

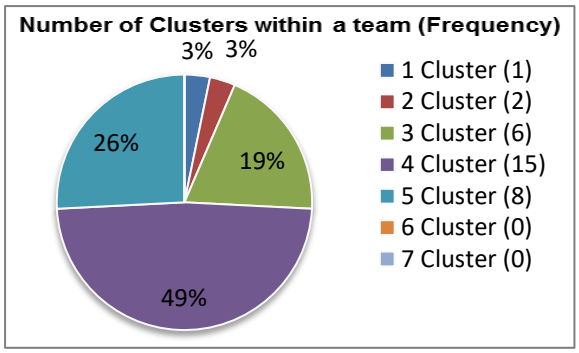


FIGURE 7.1: NUMBER OF CLUSTERS WITHIN A TEAM

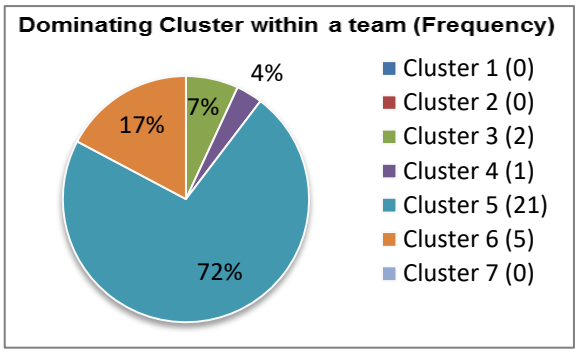


FIGURE 7.2: DOMINATING CLUSTER PER TEAM (CLUSTER WITH THE HIGHEST PROPORTION)

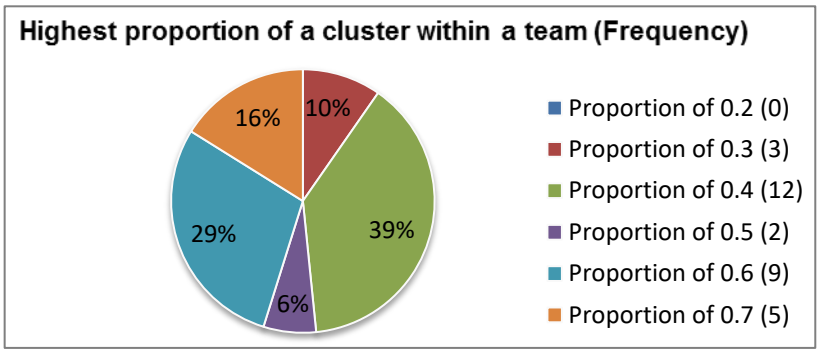


FIGURE 7.3: HIGHEST PROPORTION OF A CLUSTER WITHIN A TEAM

Layout of the Benchmarking Model

In Section 3.2 we conclude that the benchmarking model has to be clear and easy to understand. The clarity especially concerns the number of performance measure. By using the efficiency measurements defined in Section 5.2 (two measures for internal efficiency, one for the delivery and one for the depots, and one measure for external efficiency, the numbers of complaints per customer), we are able to combine different performance measures. By that we reduce the complexity of the benchmarking model, which has been one major drawback of the current benchmarking model (see Figure 3.2 – Requirement 1).

We design a new lay-out of the benchmarking model, which fulfils the requirements named in Section 3.2. It consists out of three surfaces, the further we go, the more detailed it get.

The first surface of the benchmarking model allows process managers to select the performance measure of interest. Furthermore important information can be added (“you are the top scorer”) to gain the interest of process mangers (see Figure 7.4).

On the second surface (see Figure 7.5), which can be entered by clicking on the performance measure of interest, seven column graphs are presented by default showing the top and button three delivery areas for each cluster; showing all 28 delivery areas at the beginning would only lead to confusion. If process managers want to know more about the performance measure he can click on the question mark positioned next to the heading of the surface and the performance measure record sheet of Neely et al. (1997) (see Table 4.2) for that specific performance measure will appear. This record sheet (see Table_Apx II-1) contains information about purpose, formula, data source and responsibility division of the performance measure. If process managers want to see the exact data of a column graph he can click on the “details” button next to the heading of the column graph and the measures of all influencing factors are shown. These buttons help process managers to clarify and to go into detail of performance measures if required, but still ensures compact and clear view of the performances. At the button of the second surface a graph visualises on the Dutch map the clusters and the delivery areas. Ideally, the process manager can zoom in on the map to identify

more clearly the clusters and related PC5 areas. If the process manager is interested in certain regions, delivery areas, clusters or quarters, he can select it in the field (see top left of the surface).

If the process manager wants to compare the performance measure on a lower level than the delivery area, he can click on a column of a graph and a third surface will open. Surface 3 enables a performance comparison within a delivery area (see Figure 7.6). Again each column chart will show the performance of one cluster. Thereby the process manager can choose to distinguish the performance between the different teams, depots or PC4 areas (see fields top left of Surface 2 or 3).

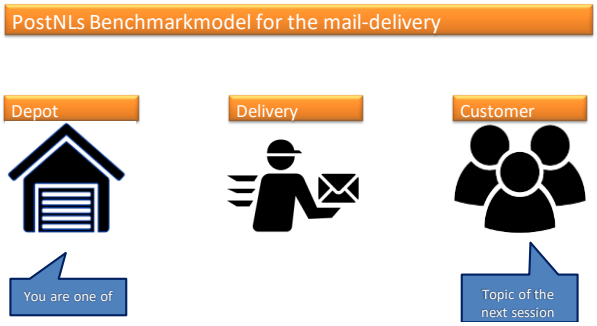


FIGURE 7.4: BENCHMARKING MODEL, SURFACE 1

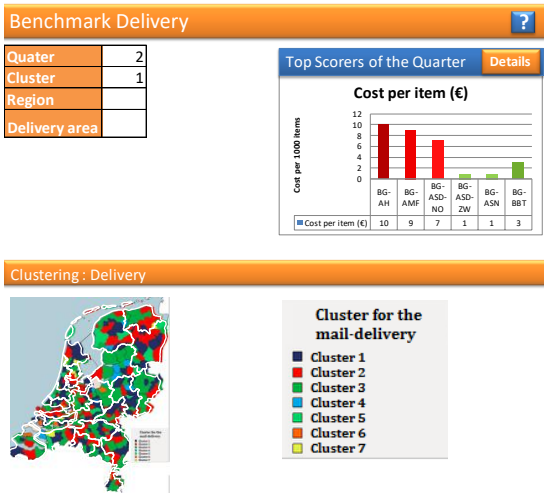
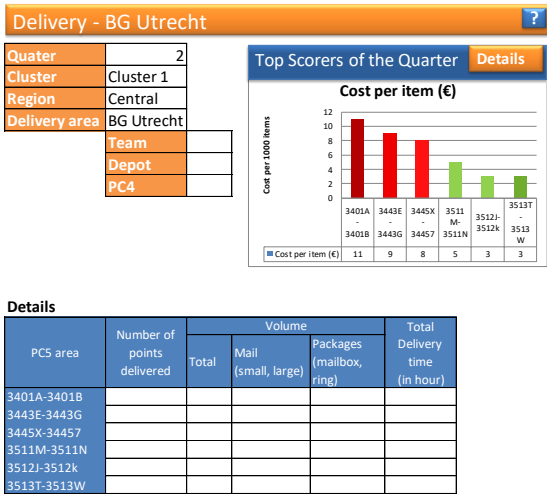


FIGURE 7.5: BENCHMARKING MODEL, SURFACE 2



Details

PC5 area	Number of points delivered	Volume			Total Delivery time (in hour)
		Total	Mail (small, large)	Packages (mailbox, ring)	
3401A-3401B					
3443E-3443G					
3445X-34457					
3511M-3511N					
3512J-3512k					
3513T-3513W					

FIGURE 7.6:BENCHMARKING MODEL, SURFACE 3

Implementation

As already discussed at the beginning of Chapter 5, we do not cover all steps which are required for the benchmarking process in this research. Steps 1, 2, 3 and 5 have already been covered in Chapters 1 to 4, which we have briefly discussed at the beginning of Chapter 5.
Steps 4, 6 and 7 are covered as follows. After applying the four perspectives framework of Neely et al. in Section 5.1, we have been able to identify costs, flexibility and quality as main critical success factors (see Figure 5.1, Step 4). With the service efficiency model of Grönroos and Ojasalo (2004) we have derived performance measures and presented the performance measure record sheet of Neely et al. to specify elements of the performance metrics (see Figure 5.1, Step 6). Finally, in Chapter 6, we have defined seven clusters for the performance measure delivery time per mail volume by using APN/km² and means of transportation as cluster attributes (see Figure 5.1, Step 7).

To ensure a successful implementation of the benchmarking model, we specify tasks, stakeholders and time required to cover the remaining steps in the following (see Table 7.1). A detailed description can be found in Appendix XV. For a general explanation of the steps we refer to Section 4.1.

For implementation PostNL still has to cover the following steps: In Step 7 PostNL has to define clusters for the remaining performance measures and collect the required data in Step 8. Subsequently, the benchmarking model has to be analysed in Step 9 in order to define performance gaps and best practices in Step 10. Subsequently, an action plan has to be defined (Step 11) to implement the best practices (Step 12). Finally, the success of the implementation should be monitored, and continuous improvement should be secured (Step 13). The exact tasks of those steps are defined in Table 7.1.

The main stakeholders required to cover those steps are the senior process manager of optimisation (O), the senior controller (C), the ambassador of delivery and the process managers (P). Looking at the duration of each step, we can see that some involve onetime tasks, for instance the clustering for the remaining performance measures (Step 7), while others are continuous like analysing and comparing once a month individually the performances (Step 9). For high time-consuming task, especially the clustering with 80 hours for each performance measure and the design of the information system with 120 hours, we recommend to set-up a small team managed by the process manager of optimisation or the senior controller to share the work.

Overall, implementing the benchmarking model will be time intensive in the short-term, however once the standards (clusters, information system, action plan template) are established, the time consume will be significant less: around 16h for a process manager, 11h for the ambassador of delivery (as he does not have to analyse each month the performances, but requires around 2h for the preparation of the benchmarking session) and 13h per quarter for the senior controller.

TABLE 7.1: IMPLEMENTATION OF THE REMAINING STEPS OF THE BENCHMARKING PROCESS (O = SENIOR PROCESS MANAGER OF OPTIMISATION, C=SENIOR CONTROLLER, A =AMBASSADOR OF DELIVERY AND P = PROCESS MANAGERS)

Step	Tasks	Person Responsible	Duration
7. Define clusters	Define clusters for remaining performance measures (PM) (see Section 5.3 and Chapter 6 as guideline)	O	80h per PM
8. Data collection	Collect for the whole Netherlands data on the cluster attributes for cluster assignment	C	16h per PM
	Design an automatic information processing system that retrieves the data and calculates the performance measures	C & IT department	120h
9. Analyse and compare	Discuss and define with process managers the expectations (e.g. time investment and tasks for benchmarking)	A	4h
	Analyse and compare	P, A, C	2h monthly (individual) 8h quarterly (plenary)
10. Define best practices and performance gaps	Summarise the findings	A, P	1h quarterly
11. Action plan	Develop an action plan template (incl. goal, task, time-framework, person responsible, methods for motoring)	O	8h
12. Implementation	Execute the action plan	P (t, pbz.)	-
13. Monitor and continuous improvement	Include the monitoring within the benchmarking model or MJ dashboards	C	5h per inclusion
	Evaluate the performance improvement and recalibrate benchmarking model if necessary	Short-term: P, A, C Long-term: C	Part of the 2h monthly analysis

8. Conclusion and Recommendation

The aim of this research is to develop an internal benchmarking model with adequate performance measures and clusters as a tool for the process managers to determine best practices and to improve the performance of national mail delivery. We defined in Section 1.2 three main problem areas, the composition of the benchmarking model with its performance indicators, the technique for clustering and the execution of the benchmarking. Our focus in this research lies on the clustering as PostNL lacks most competences and knowledge in this area.

In Section 1.7 we have defined six research questions those answers combined would fulfil our research goal. Each question is answered within one of our chapters. In Chapter 2 we have analysed the process as well as organisation of the national mail delivery and have clarified the information flow on mail delivery performance. In Chapter 3 we have conducted a stakeholder analysis using semi-structured interviews in order to evaluate the current benchmarking. By conducting an academic literature review, presented in Chapter 4, we have analysed the current scientific knowledge on benchmarking, performance measurement and cluster analysis, which we have applied to the benchmarking model of mail delivery process at PostNL in Chapter 5. In Chapter 6, we have derived a test framework by adapting clustering methods of the academic literature to PostNL. Due to the time limitation and as defined in our scope (Section 1.5), we have conducted cluster analysis based on one of the proposed performance measures. In Chapter 7 we have assessed the fit of the clustering with the managerial structure at PostNL, proposed a proto-type benchmarking model and clarified the steps for the full benchmark implementation. In this chapter we provide a short summary on our research questions (Section 8.1), but refer for detailed answers to the respective chapter. Subsequently, we name the limitations, give recommendations for PostNL (Section 8.2) and finally present topics for future research (Section 8.3).

8.1. Conclusion

1. How is the national mail delivery process of PostNL organized?

The process flow of national mail delivery of PostNL is the same everywhere in the Netherlands, qualifying a good internal benchmarking. However, there are various factors including mail volume, number of delivery points, means of transportation, distance between houses and total travel distance which differ per delivery, showing the need for clustering.

Currently, PostNL establishes an elaborated control system by implementing the so called “Manage and Justify” (MJ) dashboard for each organisational level. Those MJ dashboards are rather reactive and control tools, which do not promote learning and exchange between the different management areas. Examining the current information on mail delivery performance highlights the need for a benchmarking as well. The mail delivery time is not built on norms, but on historical data and an estimation on the volume development. There does not exist any direct information about the mail delivery performance, but it relies on information passed by the postmen or customers. The benchmarking model in contrast allows a more objective evaluation of mail delivery performance.

2. How is the current benchmarking organised?

The benchmarking consists of a benchmarking model and benchmarking sessions held 3 to 4 times a year. The goal of benchmarking is to learn from each other, which is clear for all process managers. However, during the session it has been perceived rather as a competition requiring justification than an open exchange. The process managers are willing to learn, but the current settings are not sufficient. The current benchmarking model consists of too many performance measures, and the information level is too low. Process managers find it complex and time-consuming to interpret the performance measures. The performance measures should rather give a direct indication on how well someone performs by combining different measures. Furthermore, the current clustering lacks homogeneity as the cluster objects (delivery areas) are too big and should rather be divided in multiple objects. The challenge is that many factors influence the mail delivery process, which can highly differ per delivery tour, hampering homogeneity within a cluster. Finally, the set-up of

benchmarking session is quite vague, and sessions are hold irregularly. The process managers would prefer clear guidelines and more involvement in defining the session content.

3. **How can an adequate benchmarking model for the mail delivery be designed according to the academic literature?**

For the benchmarking process there is an overall agreement on the main steps within the academic literature, which is planning and defining benchmarking elements, data collection and analysis as well as communicating and implementing best practices. Critical for a successful benchmarking is to ensure continuity as well as implementation and monitoring of the best practices.
To derive adequate performance measures different frameworks are presented within the academic literature. Applying the four perspectives on operations strategy (Slack et al., 2010) covers all relevant areas of mail delivery, and it is also combinable with other frameworks. To ensure high quality of the derived performance measures it is necessary to clarify their purpose, relation to the strategy, formula, frequency of measurement, data source, and the division of responsibility (who has to measure it and who acts on it). Therefore, we have recommend to fill-out the performance measure record sheet for each performance measure (Neely et al., 1997).
Considering cluster analysis, there is an agreement within the academic literature on the general steps, which are attribute selection and possible weighting, selection of the cluster algorithm, cluster validation and interpretation. There are many different methods presented in the literature on attribute selection/weighting and cluster algorithms. However, there is no clear framework for selecting a suitable method or algorithm. Our research contributes to the growing interest towards big data exploration for practical applications. We present different frameworks to select suitable methods for weighting attributes and clustering based on the data characteristic and information availability. By this, cluster analysis gets more accessible and provides an easier application to practical problems, which is essential due to the rise of big data and the need of companies like PostNL to explore it.

4. **Which suitable performance measures can be defined for the national mail delivery at PostNL?**

By applying the four perspectives on operations strategy (Slack et al., 2010) we conclude that costs, flexibility, quality and employee commitment are critical success factors of the mail delivery service. While the MJ dashboard already covers those factors, we propose to use performance measures on service productivity for the benchmarking model as they combine the critical success factors and allow a better comparability (relates the input towards the output). Overall, we advise to measure internal productivity (delivery time per weighted mail item and depot cost per mail item) and external productivity (complaints per delivery point) as those are essential for the success of the mail delivery service. As defined in our scope we only focus on one performance measure within the cluster analysis, which is the delivery time per weighted mail item. By evaluating impact and frequency of the parameters influencing the delivery time, we have selected interdrop, delivery points, minor-route and main-route as possible cluster attributes. Due to the low frequency, but high impact of mail-box and ring packages we will not incorporate them as cluster attributes but into the performance measure by giving each kind of mail a different weighting factor.
Finally, given the current information infrastructure of PostNL we are able to name an accurate estimator for each possible cluster attribute. As information is mostly given per post code area, we conclude that PC5 areas are the most suitable as cluster objects for our cluster analysis.

5. **What should be the clustering for the benchmarking model?**

We have tested two different scenarios, peak and off-peak, depending on the different mail volumes. Furthermore, we have applied the clustering techniques to two attribute subsets: Subset 1 including interdrop, APN/m and minor-route and Subset 2 excluding the minor-route as it does not show natural cluster tendencies with the other attributes. Performing three different clustering techniques, automatic clustering (TwoStep and x-means), clustering with K as input (k-means) and practical clustering, we have come to three conclusions. Firstly, PostNL does not have to

differentiate between peak and off-peak as objects that are within one cluster during peak days also are during off-peak. Secondly, based on the high SSE and low Silhouette Coefficient of attribute Subset 1 compared to Subset 2, we can conclude that minor-route does highly decrease the compactness of clusters and thus should be excluded from the clustering. Finally, based on the internal validation criteria (SSE and Silhouette Coefficient) k-means performs the best with a reasonable Silhouette Coefficient and the lowest SSE, closely followed by the practical technique. However, given the external validation of the expert team, the practical technique outperforms the k-means due to the good separation of means of transportation. Therefore, we select the practical approach for the benchmarking. The final cluster outcome are 7 clusters grouped by their means of transportation and a further division of the foot delivery by APN/km². Given the cluster proportion of our sample BG Utrecht, with Cluster 5 (foot delivery with APN/km² between 1000 and 3000) containing 47% of PC5 areas, we advise to re-evaluate the proportion after applying it to the whole Netherlands. If there is still one highly dominant cluster, one should split it into two equally sized clusters.
Overall, the new clustering technique which is based on PC5 areas, improves the coherence and thus similarity within a cluster by 50% compared to the original clustering which is based on delivery areas as cluster object.

6. **How should the new benchmarking model for PostNL be designed?**

PostNL has to make a trade-off between the homogeneity of clusters and the workload for extracting the required information about delivery time. As the delivery time per PC5 area does not exist, we advise to combine PC5 areas covered by the same postman into one cluster object. A postman covers 2.2 PC5 areas on average, thus chances are high that they are within the same cluster ensuring a high homogeneity. Doing this on team level would allow an easier linking to PC5 areas, however more than 84% of the teams contain PC5 areas from at least 4 different clusters, which extremely reduces the similarity within one cluster.
By designing the layout of the benchmarking model, we have demonstrated how we can keep the model clear and approachable by incorporating different surfaces with different levels of benchmarking (between delivery areas, PC4 areas or teams) and help-buttons to gain more information if required. Finally, PostNL has to cover 7 more steps to fully implement the benchmarking, which will be time consuming in the short-term, but moderate in the long-term.

8.2. Limitations

This research is based on the delivery area Utrecht with 1265 PC5 areas. Although, we have ensured that all kind of areas (from rural to highly urban, every kind of transportation means) are included, we have not assessed if the proportion is representative. This leads to two constraints: firstly theoretically, as the applied clustering techniques have been limited to non-density based algorithms. Secondly, practically, because if the final clustering regarding the whole Netherlands results in one major cluster, the clusters would not add high value to the benchmarking.

Another limitation is that we have applied theoretical clustering rather to the interdrop than to the means on transportation, which has consequently led to a low external validation. Therefore, we advise to expand our test framework by incorporating means of transportation as a nominal variable for the theoretical clustering technique. In Section 4.3.3, we have already defined the distance measure that could be used for nominal variables. Another approach is to apply the theoretical clustering techniques separately to each means of transportation.

Finally, the validation on compactness was limited to the SSE, which is favourable for k-means and x-means. To create a more fair comparison between different clustering techniques, we would advise to apply Hubert's T (see Section 4.3.5) instead, if the computational capacity is available.

8.3. Recommendation for PostNL

During our research we made several observations based on which we can formulate the following recommendations:

1. **Norming instead of Clustering**

   This research shows that there are no highly distinctive clusters and makes clear that only some attributes (interdrop and APN/km²) incorporate cluster tendencies given our performance measure delivery time per weighted mail item. Furthermore, we have shown that multiple factors influence the delivery time, which we have not incorporated in the clustering. For instance, we have been able to exclude the distance from the retailer due to the low frequency (see Section 5.3), but given the high increase in mailbox and ring-packages, we expect that it gains more importance and should be considered as well. However, the retailer has been chosen independently from the interdrop as well as the APN/km² and thus does probably not support their cluster pattern. Therefore, creating homogeneous groups for benchmarking the delivery time will be harder.

   Given that delivery has the highest expenses of the whole mail-process and that the time is not optimally calculated, there is still a high need and room for improvement, which cannot be covered by the benchmarking. To solve the problem of heterogeneity of mail delivery areas we propose to develop and to apply a norm model that allows a good estimation for the required mail delivery time. PostNL already implemented such a norm model for scooter and car deliveries, which works quite well. Furthermore, the current information infrastructure is quite elaborated. Given that in the near future exact information on the mail volume distribution is available, we expect that the required time for delivery by bike and foot can be accurately estimated. Thus, instead of the current situation where delivery time is calculated based on historical data and the volume decrease, we recommend to calculate the exact time based on the current information and the norming system.

2. **Information System**

   PostNL possesses a richness of data concerning the delivery process, which will increase in the near future due to the new machines. However, as indicated by the stakeholders (see Section 3.1), the information system contains room for improvement. Some information is available but not given on the right level like the depot costs which are only given in total for the whole region. Other valuable information could be extracted by combining data, for example, extracting the delivery time per tour by combining the delivery time of each postman with the tours he covers. Finally, some information is outdated and should be revised. In particular, information on norm values like the weighting factor 1.4 for walking stairs (see Section 5.4), especially when creating a norm model for the mail delivery time for bike and foot deliveries. Following this recommendation will help PostNL to increase the transparency of mail delivery-process and allows a better control and improvement mechanism.

3. **Crossing Managerial Boundaries**

   One major issue indicated by the stakeholders (see Section 3.1) is the link between different departments of the mail process. As defined at the beginning of this thesis, the mail process consist out of three sub-processes, collection, sorting & preparation and mail delivery process (see Figure 1.6), where the input of one process depends on the output of the other. Currently, collecting, sorting & preparation and mail delivery are managed and controlled separately. However, to ensure a smooth process flow we recommend PostNL to develop a monitoring and an evaluation tool which assesses the link between the sub-processes. In particular, this tool should compare if the output of the preceding sub-process fulfils the norms and if it is equal to the input of the succeeding sub-process.

   This will help to determine and to solve the source of problems more precisely. For instance, the current customers' complain system is only evaluated by process managers of delivery. However, some complaints can occur because of a mistake in the sorting & preparation step, therefore should be discussed together with the process manager of sorting & preparation.

4. **External Benchmarking**

   In Section 1.4 we have used the flowchart of Southard and Parente (2007) to conclude that an internal benchmarking would be beneficial. However, we recommend to develop an external benchmarking model as well, because it will expand the perspective of PostNL and help to find best global practices. US Postal Service has already implemented benchmarking as an integral part of its postal service supplying practices process (USPS, 2017). With benchmarking initiatives it was able to increase their efficiency and to lower its prices for the two-day deliveries, which enhanced its competitive composition (Yasin, 2002). This would be advisable for PostNL as well as its prices are not compatible with the ones of Sandd, its main national competitor (see Appendix VIII). The only drawback of external benchmarking is the often limited information access to competitors (Drew, 1997; Southard & Parente, 2007). However, at the logistic strategy department of PostNL we experienced that there already is some cooperation and exchange with other postal companies across the border. All of them are facing a shrinking mail market and dealing with new product, technology and strategy development to stay profitable (Chan et al., 2009). Therefore, we expect that some companies will be open for an external benchmarking.

5. **Introducing an Analytic Hierarchy Process (AHP)**

   Mail delivery is quite a complex process with various parameters and requirements. Furthermore, by analysing the critical success factors of mail delivery of PostNL from the four different operational perspectives (see Section 5.1) we saw that, depending on the perspective, different weightings of the factors are given (see Table 5.1). Similarly, the priorities of performance factors were different depending on the stakeholders (see Section 3.1). In order to incorporate such complexity within a measurement system, independent whether it is for a MJ dashboard or a benchmarking model, and still providing a clear view on the performance levels with their individual performance factors, we advise to use an Analytic Hierarchy Process (AHP) based methodology. This method places the performance in a hierarchical order, directly showing how performance measures are interlinked. Furthermore, this approach allows to give each performance measure a different weight so that different priorities can be incorporated. This AHP approach can also be applied for benchmarking (Korpela & Tuominen, 1996; Yasin, 2002). For a detailed outline see Korpela and Tuominen (1996) or Yasin (2002).

6. **Applying Data Envelopment Analysis**

   With clustering we can determine groups with high similarity. This can also be used for further benchmarking techniques. Until know we presented for each input and output factor individual performance measures (e.g. depot costs per mail item and delivery time per mail item). This enables us to benchmark specific areas. However, in order to assess the overall efficiency, we recommend to apply data envelopment analysis (DEA). DEA is a linear programming based technique which gives the relative efficiency based on multiple inputs and outputs given a certain group of decision making units (DEA) with the aim to identify efficient frontiers and best practices. For this method it is essential that the relative efficiency is measured within a group of comparable decision making units (DMUs) (Cook & Seiford, 2009), which we can ensure through our clustering. An advantage of DEA is that we can reduce the number of cluster attributes, because given that multiple inputs and outputs can be used, we can incorporate them within the performance measure for the relative efficiency (for instance the number of delivery points). Finally, this model can also be adapted to measure both the overall efficiency of multiple production stages and of the individual stages (see Zhu, 2003, cited by Cook & Seiford 2009). This can be useful for PostNL considering our Recommendation 3 as it can give one overall measure for the delivery process combined with sorting and preparation. For a general outline of DEA we refer to Boussofiane, Dyson and Thanassoulis (1991) and Cook, Tone and Zhu (2014). Additionally, Cook & Seiford (2009) present the major research thrusts in DEA which abolish some restrictions of the traditional DEA method presented by Charnes, Cooper and Rhodes (1978) (Banker, Charnes, & W.W., 1984; Cook & Seiford, 2009) and enable by that a better application.

8.4. Topics for Future Research

Our research focused on clustering techniques which are available in common statistical tools (WEKA, SPSS) to ensure easy practical application. By that we limit ourselves to 15 cluster algorithms, mostly traditional ones. Given that there are thousands of cluster algorithms available within the academic literature, some with more advanced algorithms (Halkidi, 2001; Jain, 2010), the chances are high that our framework for selecting suitable cluster algorithms (see Figure 4.10) does not provide an adequate guidance given all clustering techniques. Therefore, we suggest further research on defining a clear framework for selecting suitable clustering techniques given their characteristics on input variables and data patterns. Considering the high demand on pattern recognition within big data (Halkidi, 2001), we expect that companies are willing to expand their statistical software to incorporate the clustering techniques based on the final selection framework.

Finally, we have identified a gap within the literature of cluster validation. There are various validation criteria, however until now the majority does not clarify the threshold which has to be reached in order to be satisfying. The main reason is that it highly depends on the input, for instance the number of clusters or the range of distance measures. However, given that we can standardize the input variables to some extent (for instance standardize the attributes by the same range (Milligan & Cooper, 1988)), we believe that it is possible to develop a generally applicable table to evaluate the score of cluster criteria given a specified input (e.g. number of cluster, number of attributes or distance measure).

References

Agresti, A. (2013). *Categorical data analysis* (3rd ed.). Hoboken, NJ: Wiley.

Anand, G., & Kodali, R. (2008). Benchmarking the benchmarking models. *Benchmarking: An International Journal*, *15*(3), 257–291. https://doi.org/10.1108/14635770810876593

Anderson, K., & McAdam, R. (2004). A critique of benchmarking and performance measurement: Lead or lag? *Benchmarking: An International Journal*, *11*(5), 465–483. https://doi.org/10.1108/14635770410557708

Ankerst, M., Breunig, M. M., Kriegel, H.-P., & Sander, J. (1999). Optics: Ordering points to identify the clustering structure. *ACM Sigmod Record*, 49–60. https://doi.org/10.1145/304182.304187

Azzone, G., Masella, C., & Bertelè, U. (1991). Design of Performance Measures for Time-based Companies. *International Journal of Operations & Production Management*, *11*(3), 77–85.

Banker, R. D., Charnes, A., & W.W., C. (1984). Some Models for Estimating Technical and Scale Inefficiencies in Data Envelopment Analysis. *Managment Science*, *30*(9), 1078–1092.

Becher, J. D., Berkhin, P., & Freeman, E. (2000). Automating Exploratory Data Analysis for Efficient Data Mining. *Proceedings of the Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 424–429. https://doi.org/10.1145/347090.347179

Berkhin, P. (2006). Survey of Clustering Data Mining Techniques. *Grouping Multidimensional Data: Recent Advances in Clustering*, 25–71. https://doi.org/10.1007/3-540-28349-8_2

Binder, M., Clegg, B., & Egel-Hess, W. (2006). Achieving internal process benchmarking: guidance from BASF. *Benchmarking: An International Journal*, *13*(6), 662–687. https://doi.org/10.1108/14635770610709040

Bourne, M., & Neely, A. (2003). Implementing performance measurement systems : a literature review. *Business Performance Management*, *5*(1), 1–24.

Boussofiane, A., Dyson, R. G., & Thanassoulis, E. (1991). Applied data envelopment analysis. *European Journal of Operational Research*, *52*(1), 1–15. https://doi.org/10.1016/0377-2217(91)90331-O

Brun, M., Sima, C., Hua, J., Lowey, J., Carroll, B., Suh, E., & Dougherty, E. R. (2007). Model-based evaluation of clustering validation measures. *Pattern Recognition*, *40*(3), 807–824. https://doi.org/10.1016/j.patcog.2006.06.026

Brusco, M. J., & Cradit, J. D. (2001). A variable-selection heuristic for K-means clustering. *Psychometrika*, *66*(2), 249–270. https://doi.org/10.1007/BF02294838

Burnham, K. P., & Anderson, R. P. (2004). Multimodel Inference: Understanding AIC and BIC in Model Selection. *Sociological Methods & Research*, *33*(2), 261–304. https://doi.org/10.1177/0049124104268644

Camp, R. C. (1989). *Benchmarking: The Search for Industry Best Practices that Lead to Superior Performance* (1st ed.). Milwaukee, Wis: Quality Press.

Celeux, G., & Govaert, G. (1992). A classification EM algorithm for clustering and two stochastic versions. *Computational Statistics & Data Analysis*, *14*(3), 315–332. https://doi.org/10.1016/0167-9473(92)90042-E

Chan, F. T. S., Henry, H. K. C., & Ralph, C. W. L. (2009). An AHP approach in benchmarking logistics performance of the postal industry. https://doi.org/10.1108/14635770610709031

Cook, W. D., & Seiford, L. M. (2009). Data envelopment analysis (DEA) - Thirty years on. *European Journal of Operational Research*, *192*(1), 1–17. https://doi.org/10.1016/j.ejor.2008.01.032

Cook, W. D., Tone, K., & Zhu, J. (2014). Data envelopment analysis: Prior to choosing a model. *Omega (United Kingdom)*, *44*, 1–4. https://doi.org/10.1016/j.omega.2013.09.004

Dattakumar, R., & Jagadeesh, R. (2003). *A review of literature on benchmarking*. *Benchmarking: An International Journal* (Vol. 10). https://doi.org/10.1108/14635770310477744

De Veaux, R. D., Velleman, P. F., & Bock, D. E. (2012). *Stats: Data and Models* (3rd ed.). Boston: Pearson, Addison-Wesley.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society Series B Methodological*, *39*(1), 1–38.

https://doi.org/http://dx.doi.org/10.2307/2984875
Drew, S. a. W. (1997). From knowledge to action: the impact of benchmarking on organizational performance. *Long Range Planning*, *30*(June), 427–441. https://doi.org/10.1016/S0024-6301(97)90262-4
Dubes, R. C. (1993). Cluster analysis and related issues. In C. H. Chen, L.-F. Pau, & P. S.-P. Wang (Eds.), *Handbook of pattern recognition & computer vision* (pp. 3–32). River Edge, NJ, USA: World Scientific Publishing Co.
Fortuin, L. (1988). Performance indicators – why, where and how? *European Journal of Operational Research*, *34*(1), 1–9.
Frey, B. J., & Dueck, D. (2007). Clustering by passing messages between data points. *Science (New York, N.Y.)*, *315*(5814), 972–976. https://doi.org/10.1126/science.1136800
Gnanadesikan, R., Kettenring, J. R., & Tsao, S. L. (1995). Weighting and selection of variables for cluster analysis. *Journal of Classification*, *12*(1), 113–136. https://doi.org/10.1007/BF01202271
Goold, M. (1991). Strategic control in the decentralised firm. *Sloan Management Review*, *32*(2), 69–81.
Goold, M., & Quinn, J. J. (1990). The paradox of strategic controls. *Strategic Management Journal*, *11*, 43–57.
Gremler, D. D. (2004). The Critical Incident Technique in Service Research. *Journal of Service Research*, *7*(1), 65–89. https://doi.org/10.1177/1094670504266138
Grönroos, C., & Ojasalo, K. (2004). Service productivity - Towards a conceptualization of the transformation of inputs into economic results in services. *Journal of Business Research*, *57*(4), 414–423. https://doi.org/10.1016/S0148-2963(02)00275-8
Halkidi, M. (2001). On Clustering Validation Techniques. *Journal of Intelligent Information Systems*, *173*(2), 107–145.
Halkidi, M., Batistakis, Y., & Vazirgiannis, M. (2002a). Cluster Validity Methods : Part I, *31*(2), 40–45.
Halkidi, M., Batistakis, Y., & Vazirgiannis, M. (2002b). Clustering Validity Methods: Part II. *ACM SIGMOD Record*, *31*(3), 19–27.
Hall, L. (1996). Computational Complexity. In S. I. Gass & C. M. Harris (Eds.), *Encyclopedia of Operations Research and Management Science* (pp. 95–98). Boston, MA: Springer US. https://doi.org/10.1007/978-1-4613-0459-3
Hall, M. A., & Holmes, G. (2003). Benchmarking Attribute Selection Techniques for Discrete Class Data Mining. *IEEE Transactions on Knowledge and Data Engineering*, *15*(6), 1437–1447. https://doi.org/10.1109/TKDE.2003.1245283
Heerkens, J. M. G. (2004). *A Methodological Checklist for the High-Tech Marketing Project*. *TSM Business School*.
Huang, J. Z., Ng, M. K., Rong, H., & Li, Z. (2005). Automated variable weighting in k-means type clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *27*(5), 657–668. https://doi.org/10.1109/TPAMI.2005.95
IBM. (2016). IBM SPSS Statistics 24 Algorithms. Retrieved from http://www-01.ibm.com/software/analytics/spss/
Jain, A. K. (2010). Data clustering: 50 years beyond K-means. *Pattern Recognition Letters*, *31*(8), 651–666. https://doi.org/10.1016/j.patrec.2009.09.011
Jain, A. K., & Dubes, R. C. (1988). Algorithms for Clustering Data. *Prentice Hall*. https://doi.org/10.1126/science.311.5762.765
Jain, A. K., Duin, R. P. W., & Mao, J. (2000). Statistical pattern recognition: a review. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *22*(1), 4–37. https://doi.org/10.1109/34.824819
Jain, A. K., Murty, M. N., & Fylnn, P. J. (2000). Data Clustering : A Review, *1*(212).
Jing, L., Ng, M. K., & Huang, J. Z. (2007). An entropy weighting k-means algorithm for subspace clustering of high-dimensional sparse data. *IEEE Transactions on Knowledge and Data Engineering*, *19*(8), 1026–1041. https://doi.org/10.1109/TKDE.2007.1048

Johnson, G., Whittington, R., & Scholes, K. (2011). *Exploring Strategy* (9th ed.). Harlow, England: Pearson Education Limited.

Kaplan, R. S., & Norton, D. P. (1992). The Balanced Scorecard - Masures That Drive Performance. *Harvard Business Review*, *70*(1), 71–79. https://doi.org/00178012

Ketchen, D. J., & Shook, C. L. (1996). The Application of Cluster Analysis in Strategic Management Research: An Analysis and Critique. *Strategic Management Journal*, *17*(6), 441–458. https://doi.org/10.1002/(SICI)1097-0266(199606)17:6<441::AID-SMJ819>3.0.CO;2-G

Kettenring, J. R. (2006). The practice of cluster analysis. *Journal of Classification*, *23*(1), 3–30. https://doi.org/10.1007/s00357-006-0002-6

Korpela, J., & Tuominen, M. (1996). Benchmarking logistics performance with an application of the analytic hierarchy process. *IEEE Transactions on Engineering Management*, *43*(3), 323–333. https://doi.org/10.1109/17.511842

Kriegel, H.-P., Kröger, P., & Zimek, A. (2009). Clustering high-dimensional data. *ACM Transactions on Knowledge Discovery from Data*, *3*(1), 1–58. https://doi.org/10.1145/1497577.1497578

Kuha, J. (2004). AIC and BIC: Comparisons of Assumptions and Performance. *Sociological Methods & Research*, *33*(2), 188–229. https://doi.org/10.1177/0049124103262065

Landeghem, R. Van, & Persoons, K. (2001). Benchmarking of logistical operations based on a causal model. *International Journal of Operations & Production Management*, *21*(1), 254–267. https://doi.org/10.1108/01443570110358576

Larsen, R. J., & Marx, M. L. (2012). *An Introduction to Mathematical Statistics and Its Applications* (5th ed.). Prentice Hall.

Law, M. H. C., Figueiredo, M. A. T., & Jain, A. K. (2004). Simultaneous feature selection and clustering using mixture models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *26*(9), 1154–1166. https://doi.org/10.1109/TPAMI.2004.71

Lea, R., & Parker, B. (1989). The JIT spiral of continuous improvement. *IMDS*, *4*, 2025–2036.

Lynch, R. L., & Cross, K. F. (1991). *Measure up – The Essential Guide to Measuring Business Performance*. London: Mandarin.

Maskell, B. H. (1991). *Performance Measurement for World Class Manufacturing*. Cambridge, MA: Productivity Press.

Milligan, G. W., & Cooper, M. C. (1988). A study of standardization of variables in cluster analysis. *Journal of Classification*, *5*(2), 181–204. https://doi.org/10.1007/BF01897163

Neely, A., Gregory, M., & Platts, K. (1995). Performance measurement system design: A Literature Review and Research Agenda. *International Journal of Operations & Production Management*, *15*(4), 80–116. https://doi.org/http://dx.doi.org/10.1108/01443579510083622

Neely, A., Mills, J., Platts, K., Richards, H., Gregory, M., Bourne, M., & Kennerley, M. (2000). Performance measurement system design: developing and testing a process-based approach. *International Journal of Operations & Production Management*, *20*(10), 1119–1145. https://doi.org/10.1108/01443570010343708

Neely, A., Richards, H., Mills, J., Platts, K., & Bourne, M. (1997). Designing performance measures: a structured approach. *International Journal of Operations & Production Management*, *17*(11), 1131–1152. https://doi.org/10.1108/01443579710177888

Pelleg, D., & Moore, A. W. (2000). X-means: Extending K-means with efficient estimation of the number of clusters. *Proceedings of the Seventeenth International Conference on Machine Learning Table of Contents*, 727–734. https://doi.org/10.1007/3-540-44491-2_3

Raftery, A. E., & Dean, N. (2006). Variable Selection for Model-Based Clustering. *Journal of the American Statistical Association*, *101*(473), 168–178. https://doi.org/10.1198/016214506000000113

Ray, G., Barney, J. B., & Muhanna, W. A. (2004). Capabilities, business processes, and competitive advantage: Choosing the dependent variable in empirical tests of the resource-based view. *Strategic Management Journal*, *25*(1), 23–37. https://doi.org/10.1002/smj.366

Reed, M. S., Graves, A., Dandy, N., Posthumus, H., Hubacek, K., Morris, J., … Stringer, L. C. (2009).

Who's in and why? A typology of stakeholder analysis methods for natural resource management. *Journal of Environmental Management*, *90*(5), 1933–1949. https://doi.org/10.1016/j.jenvman.2009.01.001

Rockert, J. F. (1979). Chief executive define their own data needs. *Harvard Business Review*. https://doi.org/Article

Rousseeuw, P. J. (1987). Silhouettes-a graphical aid to the interpretation and_validation of cluster analysis, *20*, 53–65. https://doi.org/10.1016/0377-0427(87)90125-7

Slack, N., Chambers, S., & Johnston, R. (2010). *Operations Management* (6th ed.). Harlow, England: Financial Times Prentice Hall.

Southard, P. B., & Parente, D. H. (2007). A model for internal benchmarking: when and how? *Benchmarking: An International Journal* , *14*(2), 161–171. https://doi.org/10.1108/14635770710740369

Spendolini, M. J. (1992). *The benchmarking book* (1st ed.). New York: Amacom.

SPSS. (2001). The SPSS TwoStep Cluster Component: A scalable component enabling more efficient customer segmentation, 9. Retrieved from http://www.spss.ch/upload/1122644952_The SPSS TwoStep Cluster Component.pdf

Steinley, D. (2003). Local optima in K-means clustering: what you don't know may hurt you. *Psychological Methods*, *8*(3), 294–304. https://doi.org/10.1037/1082-989X.8.3.294

Steinley, D. (2006). K-means clustering: a half-century synthesis. *The British Journal of Mathematical and Statistical Psychology*, *59*(Pt 1), 1–34. https://doi.org/10.1348/000711005X48266

Steinley, D., & Brusco, M. J. (2008a). A New Variable Weighting and Selection Procedure for K - means Cluster Analysis. *Multivariate Behavioral Research*, *43*(1), 77–108. https://doi.org/10.1080/00273170701836695

Steinley, D., & Brusco, M. J. (2008b). Selection of variables in cluster analysis: An empirical comparison of eight procedures. *Psychometrika*, *73*(1), 125–144. https://doi.org/10.1007/s11336-007-9019-y

Subramanian, N., & Ramanathan, R. (2012). A review of applications of Analytic Hierarchy Process in operations management. *International Journal of Production Economics*, *138*(2), 215–241. https://doi.org/10.1016/j.ijpe.2012.03.036

Tan, P.-N., Steinbach, M., & Kumar, V. (2005a). Classification : Basic Concepts , Decision Trees , and. In *Introduction to Data Mining* (1st ed., pp. 145–206). Pearson. https://doi.org/10.1016/0022-4405(81)90007-8

Tan, P.-N., Steinbach, M., & Kumar, V. (2005b). Cluster Analysis: Basic Concepts and Algorithms. In *Introduction to Data Mining* (1st ed., pp. 487–568). Pearson. https://doi.org/10.1016/0022-4405(81)90007-8

Tibshirani, R., Walther, G., & Hastie, T. (2001). Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *63*, 411–423. https://doi.org/10.1111/1467-9868.00293

Voss, A., Åhlström, P., & Blackmon, K. (1997). Benchmarking and operational performance : some empirical. *International Journal of Operations & Production Management*, *17*(10), 1046–1058.

Xu, R., & Wunsch, D. (2005). Survey of Clustering Algorithms. *IEEE Transactions on Neural Networks*, *16*(3), 645–678. https://doi.org/10.1109/TNN.2005.845141

Yasin, M. (2002). *The theory and practice of benchmarking: then and now*. *Benchmarking: An International Journal* (Vol. 9). https://doi.org/10.1108/14635770210428992

Web links:

ACM (2016, December 05). Ontwikkelingen en feiten bij de Nederlandse postmark in 2015. Retrieved from ACM Website: https://www.acm.nl/nl/publicaties/publicatie/16592/Ontwikkelingen-en-feiten-bij-de-Nederlandse-postmarkt-in-2015/

PostNL. (2016a, October 17). Annual Report 2015. Retrieved from PostNL Web site: http://www.postnl.nl/over-postnl/over-ons/jaarverslag/

PostNL. (2016b, October 17). Interactive charts. Retrieved from PostNL Web site: http://annualreport2015.postnl.nl/docs/PostNL_AR2015/interactive-charts.php?pst=sp-pnl-pos-over-postnlover-onsjaarverslag-footerbanner3-interactive_charts

PostNL (2016c, November 29). Mail in the Netherlands: keeping up the pace. Retrieved from PostNL Web site: https://www.postnl.nl/en/about-postnl/about-us/our-stories/#trigger-75018

PostNL. (2016d, October 17). Mail in the Netherlands. Retrieved from PostNL Web site: http://www.postnl.nl/en/about-postnl/about-us/our-organisation/mail-in-the-netherlands/

PostNL. (2016e, October 17). The history of mail. Retrieved from PostNL Web site: http://www.postnl.nl/en/about-postnl/about-u istory/s/h

Sandd. (2016a, December 05). Klantenservice. Retrieved from Sandd Web site: http://www.sandd.nl/klantenservice/

Sandd. (2016b, December 05). Wij zijn Sandd. Retrieved from Sandd Web site: http://www.sandd.nl/over-sandd/

USPS. (2017, May 10). Supplying Principles and Practices Manual. Retrieved from United States Postal Service Web site: https://about.usps.com/manuals/spp/html/welcome

Appendices

Appendix I Mail Delivery Process

In the following we describe the steps of the mail delivery process. Therefore, we first need to know how the mail is sorted, secondly what steps the postman has to do at the depot or HUB before starting the delivery tour, thirdly how the actual delivery is done and finally the task and responsibilities of the postman to close up the process.

Sorting of the Mail

The postman receives the mail at the depot or HUB already sorted and prepared in inner bags at the depot. The inner bags have a standard size which fits perfectly in the bags of the postman. Unaddressed mail, mailbox packages and the remaining addressed mail are separated into different inner bags; the main reason lies in the sorting & preparation process: unaddressed mail does not have to be sorted and hence can be directly placed into the inner bag. Addressed mail is sorted based on the street number and the mail delivery route. The sorted mail is bundled with elastic bands based on the sub tours within the mail delivery tour of the postman. Those bundles are placed on order in the inner bags, so that the upper bundle is always the next in sequence for the delivery. Hence, the postman does not have to sort the mail on the street, but can just grab the upper bundle.

To prohibit damages on the mailbox packages due to the weight of the mail, the mailbox packages are not included in the bundle, but placed in a separate inner bag. Each bundle that should include a mailbox packages gets a pink card reminding the postman to grab the mailbox package out of the other inner bag.

Overall, the postman receives inner bags with three different fillings. To enable a better differentiation the inner bags with unaddressed mail are grey and with addressed mail orange. Furthermore, each orange inner bag is tacked with an identification card which shows the depot and delivery tour number.

Preparation at the HUB/ Depot

The mail delivery process is every Tuesday, Wednesday, Thursday, Friday and Saturday and can be done by foot, bike, e-bike, scooter or car (see Appendix II). Each delivery tour is assigned to one certain depot, except the delivery tours by cars, those pick up the mail at the HUBs. Mostly the delivery tour is assigned to the closest depot or HUB to minimise extra traveling (distance from deport to main delivery), called run-up.

The mail to the HUBs is delivered before 9 a.m., because postmen that delivery by car, have to start their tour at 9.30 a.m. For car deliveries the inner bags stay at the roll container after the arrival at the HUB. All the inner bags of one roll container are for the same delivery tour. The postman moves the container in front of his car and places the inner bags directly from the roll container in the car. Afterwards the postmen can depart with the car.

Mail to the depot is delivered at different time slots, one at 11 a.m. and the other at 1 p.m. Deliveries from there have no mandatory starting time, but do have to be finished before 6 p.m.

The quantity of inner bags is for foot, scooter, bike and e-bike delivery significant less than for the car. Consequently, one roll container often contains inner bags of different delivery tours. Therefore, after the roll containers arrive at the depot, one postman is assigned to sort all the inner bags by placing them in the shelf with the same delivery tour number.

The postman arrives by foot, with his own bike, a leased PostNL scooter or a PostNL e-bike at the depot. Every postman borrows his own equipment (PostNL clothes and bags) from PostNL and stores it at home, except for the bag carrier for foot deliveries, the so called post-boy, and the bike trailer (see Appendix II), which are stored at the depot.

After the arrival the postman controls his/her bags for remaining items from the previous delivery tour, which can includes empty inner bags, elastic bands or garbage, and sort them out to the corresponding places of the depot. Subsequently the postman takes the inner bags of his/her

delivery tour from the shelf to his/her means of transportation and places them in the main bags. If there are more inner bags than that can fit on the means of transportation, the postman can try to spread the mailbox packages or unaddressed mail by taking them out of the inner bag and placing them directly in the main bags. However, if that does not fit, they have to split the mail into two and have to come back during their delivery tour picking up the remaining mail. Afterwards the postman can go with his/her mean of transportation to the starting point of the mail-deliver route (see Figure 2.1, no. 1).

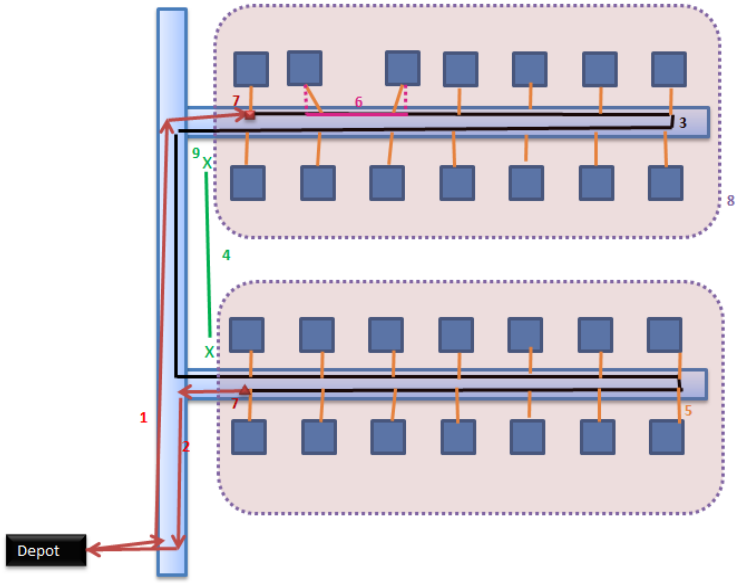
Mail delivery

Every postman has a certain route for his delivery tour (see Figure 2.1, no.3), that he has to follow, called main-route. The length of the main-route varies per delivery tour, but has a limitation per means of transportation.

Some delivery tours, mostly the ones by bike, contain sub-tours (see Figure 2.1, no. 8) , where the postman has to park (see Figure 2.1, no. 9) and step off his/her means of transportation, take the bundle, unaddressed mail and mailbox packages for the sub-tour out of the bag and walks one round for the mail delivery. If the delivery tour has no sub-tours, the postman can stay with his means of transportation. The sequence on delivering the addresses is for all tours specified.

The mailboxes are not always reachable from the street. Often, for instance if houses have front yards, the postman have to walk a minor-route from the street to the mailbox (see Figure 2.1, no. 5). During the walking the postman grabs the addressed mail out of the bundle and if necessary mailbox packages and unaddressed mail. If it is not possible to put all the mail in the mailbox, the postman rings the bell of that address and tries to hand it in in person. If no one opens, the postman tries to contact the neighbours so that they can forward the mail later on. Otherwise the postman will bring the mail at the end of his/her delivery-tour to a certain retailer. To inform the address-holder, the postman has a standardised form informing about the location of the post (neighbours or retailer) which he/she fills in and subsequently puts it into the mailbox of the householder. Furthermore, he puts a sticker on the not deliverable mail providing information for the retailer about the householder and delivery day.

If the sub-tour is finished, the postman goes back to his/her means of transportation and rides/drives to the next delivery point or to the parking spot of the next sub-tour until he/she reaches the end of the main-route.



No.	Term
1	run-up
2	run-off
3	main-route
4	connection route
5	minor-route
6	interdrop
7	start- and endpoint
8	sub-tour
9	parking spot

FIGURE_APX I-1: MAIL DELIVERY TOUR

It can be that some mail could not be delivered. This can be due to three reasons. Firstly, because the mail is sorted in the wrong tour, meaning that the address is not within that certain delivery tour, but in another. Secondly, because the address does no longer exist or the mail is not accepted by the householder. Finally, because the mail, in this case mostly mailbox packages, does not fit through the mailbox. All this mail has to be equipped with a sticker informing about the cause. If possible, the mail has to be put by the postman into the public mailbox of PostNL. Otherwise the postman has to bring it to a certain retailer.
After taking care of the undeliverable mail, the postman can drive directly home, except if they have to return the post boy or the bike trailer to the depot.
For deliveries by car this process is slightly different. The postman does not have to drive to a retailer or a mailbox of PostNL, because they have to return the car to the HUB and at the HUB are boxes for the undeliverable mail. Before going home, they have to place all the empty inner bags and elastic bands on specific place in the HUB.

Appendix II Means of Transportation

PostNL uses different kind of means of transportation for the mail delivery. The most common once are the postboy, the (e-) bike, the bike trailer, scooter and car.

FIGURE_APX II-1: POST BOY

FIGURE_APX II-2: BIKE

FIGURE_APX II-3: BIKE TRAILER

FIGURE_APX II-4: CAR

FIGURE_APX II-5: SCOOTER

Appendix III Stakeholder Positioning

For the stakeholder analysis different types of frameworks are available. Frameworks that are used to identify the stakeholders, to categorise them or to investigate the relationship between the stakeholder (Reed et al., 2009). To classify stakeholders there are two different approaches. One is the analytical categorisation, where the stakeholder classification is conducted based on the observations and knowledge of the system of those who carry out the analysis. The other, the so called reconstructive categorisation, follows a bottom-up approach, where the stakeholder themselves define categories which provides more insight into their concerns. (Reed et al., 2009) As we already conducted many observations and an analysis of the mail delivery process as well as the management system, it will be more efficient and less time consuming to use the analytical approach.

One of the most used methods within the analytical categorisation is the power-interest matrix. Stakeholders are plotted into that matrix depending on their degree of interest and influence. This helps specifying to what extent a stakeholder wants and should be engaged in the regarding project. (Reed et al., 2009)

In the following we analyse for each stakeholder (postmen, process managers, ambassador of the delivery, senior controller and the senior process manager of optimisation) their position in the power-interest grid (Slack et al., 2010) for the mail delivery benchmarking model (see Figure 3.1) and define their degree of involvement in the benchmarking.

The **postman** is solely focused on his/her own delivery tour. The postman has no active say in the mail delivery process as the routing and the time for one tour is predetermined. Only once or twice a year, when the time for a tour is changed by the process optimisation department, the postman gets the chance to negotiate. Overall, we conclude that the interest as well as the power of the postmen is low. Thus we do not involve the postmen in the benchmarking, but as they are the last chain of the mail delivery process, we will interview them briefly to get a deeper understanding of the actual mail delivery.

Each **team leader** is responsible to manage the postmen of certain delivery tours. Their area of responsibility starts at the depot/ HUB and ends at the customer, thus exactly the scope of this project. They have direct contact with not only the postmen, but also with the customers by visiting them to deal with their complains or assessing their satisfaction with the service of PostNL. However, they do not have the power to influence or change the delivery process on their own. They can only give suggestions to their process manager, who makes the final decision which the team leader has to follow. Thus they are interested, but have low power. Therefore, we should keep them informed over the results of the benchmarking, but not actively involve them. Until now, it is working well as the process manager passes important and relevant findings, which are made during the benchmarking sessions, to the team leaders.

Each **process manager** is responsible for the delivery area and has to manage and justify the performance through the MJ dashboards. They have the power to change or adapt the mail delivery process (e.g. location of the depots, numbers of postmen) if he sees the urgency. However, radical changes need to be accepted by the region manager or other higher departments. Hence, process managers have medium till high power. The benchmarking is a comparison on a process manager level and hence should be of their interest. The process manager is interested in the performance of their colleagues and keen to learn, therefore we expect high interest. However, their major priority lies on performing well according to the performance indicators of the MJ dashboards as they are evaluated by the region manager mainly based on that outcome. Hence, they do want a good benchmarking model and session, but rather spend as little time as possible on it. Thus, they have medium interest. Therefore, according to the power-interest matrix, it is important to keep them satisfied, but only involve them if they gain value out of it. To ensure that they are satisfied with the

benchmarking model, we do interview a representative sample of process managers to determine their interests and needs.
The **control department** is among others responsible for keeping an overview of the performance of the production department. They assess the realisation versus the budget for the MJ dashboards to determine if someone is underperforming and should improve. The benchmarking model, in contrast, has no fixed targets or budgets as the aim is to keep an open space for discussion without right or wrong and rather stimulate the learning than forcing to reach a certain target. Within the control department, there is one senior controller who is responsible for the project related to delivery and hence deals with the benchmarking model and sessions. She did not determine the model, but has the power to make adaptions if necessary. Her responsibility is to gather the data and summarise the performance results and abnormalities of the benchmarking model per quarter in a PowerPoint presentation for the process managers. By determining the content of the presentation she has direct influence on the topics that are discussed during the benchmarking session. Hence, overall she is quite powerful.
Every change in the model or session lies within her responsibility. Thus, she is interested in everything that concerns the benchmarking.
Overall, we should highly involve her and ensure her satisfaction as without her agreement, we could not make any changes.
While the senior control is doing the groundwork by providing the data, the **ambassador of delivery**, who is one of the three region manager, is responsible for the execution of the benchmarking sessions. He has to ensure that the session proceeds well by leading them and that the process managers are satisfied and can gain some value out of it. He has an active say in the structure and management of the benchmarking sessions. Therefore, he is a powerful stakeholder. For him it is important that the process managers can improve through the benchmarking model and he tries to support that as much as possible. Hence, he is not only a powerful, but also an interested stakeholder. Thus, we should ensure that he is satisfied and engage hem especially in areas related to the execution of the benchmarking.
The last stakeholder is the **senior process manager of optimisation**. He is responsible for improving the overall quality and the process of the mail delivery. As the benchmarking has the goal of learning from each other to improve the delivery process, he is attending the sessions as well. His interest in developing an adequate benchmarking model is quite high as it can be good tool for improving the mail delivery process. However his responsible is rather the overall improvement of the mail delivery process, than focussing on the individual level of a process manager. Concerning the benchmarking model or sessions he has no active role in design, but more an advisor role as an expert.
Overall, his interest in the benchmarking is high, but his power is low. Thus, we inform him over the decisions and ask for his advice, but it is not critical to satisfy him.

Appendix IV Interview Templates

The stakeholders were directly contacted via phone or email with the request for an interview. We conducted the interviews face-to-face which we audio recorded and wrote down in notes (for the less powerful stakeholders) and in a transcript (for the powerful stakeholder) for further analysis. The interviews were taken in a time period of two weeks. We designed two interview templates, one for stakeholders who are joining the benchmarking and one for the others. Those two templates are nearly identical, only the questions about the current benchmarking model and sessions are replaced by general questions about benchmarking the mail delivery process. The interviews start with a critical incident question to explore how the current learning and improving is within the mail delivery process. This critical incident technique gives the respondent the opportunity to select one specific event from their own experience. It shows his/her way of thinking as we do not use a certain framework or restrict it to specific variables or actions. By this we can see what is important for the respondent (Gremler, 2004). Furthermore, learning and improving can happen without the awareness of the respondent and recalling a specific events makes it easier for the respondent to answer our questions. Subsequently, we use semi-structured interview questions on performance evaluation, clustering and benchmarking as we are not certain yet of the interest and needs of the stakeholder, but still want to enable comparison between the interviews (see Section 1.7 for a more detailed explanation).

INTERVIEW PROTOCOL TEAM LEADER/POSTMEN

- ask if you can record the interview-

Introduction of yourself and the project

First we introduce ourselves and thank the respondent for taking the time to be interviewed

We explain briefly why we are doing this study:

a) we would like to improve the current mail delivery process
b) learn more about the mail delivery process, in particular on performance evaluation

Introductory information on the respondent's background

- Name of respondent(Naam van de respondent)
- Name of function / position in the organisation / main task-responsibility (Functie naam, positie in de organisatie, voornaamste taak-veranwoordelijkheid)
- Experience in this specific position (Ervaring in deze specifieke positie)
- Total work experience at PostNL (Totale ervaring bij PostNL)

Interview question (critical incident technique)

1. Could you mention an example when you learnt something valuable for improving the mail delivery process?
    - In which situation was that? What was the initiation and motivation?
    - What did you or others particularly do?
    - How did you apply it to the mail delivery process? Could it be applied universally to delivery tours?

Kan je een voorbeeld geven waarin je iets waardevols voor de verbetering van het bezorgproces hebt geleerd?

- In wat voor een situatie was het? Waar ging het over? Wat was de uitgangspunt/ initiatie?
- Kunt u in meer detail beschrijven wat je of iemand anders precies deed?
- Hoe heb je het geleerde op het bezorgproces toegepast/ geïmplementeerd? Kon je het overal op dezelfde manier toepassen?

Stakeholder needs

2. What do you think is working well in the mail delivery process? What do you think requires improvements?
3. What (tool) do you think is useful to improve the mail delivery process?
4. What do you think are the most important (non-financial) aspects of mail delivery performance?
5. Are those factors equally weighted or do you have priorities?

- Wat werkt goed in het bezorg proces? Wat kan worden verbeterd?
- Wat zou handig zijn om de bezorging te verbeteren?
- Wat zijn volgens jou de meest belangrijke aspecten binnen de bezorging?
- Zijn de aspecten even belangrijk of hoe zou je die ranken?

Performance evaluation:

"A performance measure can be defined as a metric used to quantify the efficiency and/or effectiveness of an action" (Neely et al., 1995, p. 1229)

6. How is your performance currently evaluated?
    a. What do you think are advantages and disadvantages of this way of evaluation?
    b. How would you change the way of evaluation if you could?
7. When is in your opinion performance evaluation useful?
8. Based on which factors would you evaluate the mail delivery performance?
9. How and when would you measure those factors?
10. Do you have any performance evaluation tools next to the official tools given by the control department?

- Op wat voor een manier wordt jouw performance geëvalueerd?
    - Wat zijn volgens jou de sterkte en zwakte punten?
    - Hoe zou je het willen veranderen?
- Wanneer is volgens jou performance evaluatie nuttig?
- Met welk factoren zou je de performance van de bezorging meten?
- Hoe en wanneer zou je deze factoren meten?
- Gebruik je naast de dashboards en het benchmarking model van controlling nog andere hulpmiddel om de bezorg performance te evalueren/ controleren

Benchmarking model

11. What do you think are important elements if you want to compare the performance of one delivery tour with another? What factors would you compare?
12. How would you ensure a fair comparison?

- Welke elementen zijn volgens jou belangrijk als je de performance van een bezorg loop met andere bezorg lopen wil vergelijken? Welk factoren zou je willen vergelijken?
- Hoe kan je een faire vergelijking creëren?

Clustering:

13. What is in your opinion a perfect clustering for the mail delivery process?
14. Which aspects would you incorporate in the clustering?
    - Wat is de perfecte clustering voor het bezorgen volgens jou?
    - Welke aspecten zou je voor de clustering gebruiken?

Closure of the interview

15. Do you have any final comments or thoughts on this matter you would like to share? Heb je nog andere opmerkingen of gedachten die je zou willen delen?

INTERVIEW PROTOCOL MANAGEMENT

- ask if you can record the interview –

Introduction of yourself and the project

First we introduce ourselves and thank the respondent for taking the time to be interviewed

We explain briefly why we are doing this study:

c) we would like to improve the current mail delivery process
d) learn more about the mail delivery process, in particular on performance evaluation

Introductory information on the respondent's background

- Name of respondent(Naam van de respondent)
- Name of function / position in the organisation / main task-responsibility (Functie naam, positie in de organisatie, voornaamste taak-veranwoordelijkheid)
- Experience in this specific position (Ervaring in deze specifieke positie)
- Number of direct reports (=people that directly report to the you in the formal hierarchy of the organisation) (Schatting #medewerkers die direct onder u vallen)
- What type of work do people under hem/her (direct reports and others in the hierarchy below manager) (Type werk van medewerkers onder u in de hiërarchie)
- Total work experience at PostNL (Totale ervaring bij PostNL)

Interview question (critical incident technique)

16. Could you mention an example when you learnt something valuable for improving the mail delivery process?
    a. In which situation was that? What was the initiation and motivation?
    b. What did you or others particularly do?
    c. How did you apply it to the mail delivery process? Could it be applied across the whole delivery area equally?

Kan je een voorbeeld geven waarin je iets waardevols voor de verbetering van het bezorgproces hebt geleerd ?

- In wat voor een situatie was het? Waar ging het over? Wat was de uitgangspunt/ initiatie?
- Kunt u in meer detail beschrijven wat je of iemand anders precies deed?
- Hoe heb je het geleerde op het bezorgproces toegepast/ geïmplementeerd? Kon je het overal op dezelfde manier toepassen?

Stakeholder needs

17. What do you think is working well in the mail delivery process? What do you think requires improvements?
18. What (tool) do you think is useful to improve the mail delivery process?
19. What do you think are the most important (non-financial) aspects of mail delivery performance?
20. Are those factors equally weighted or do you have priorities?

- Wat werkt goed in het bezorg proces? Wat kan worden verbeterd?
- Wat zou handig zijn om de bezorging te verbeteren?
- Wat zijn volgens jou de meest belangrijke aspecten binnen de bezorging?
- Zijn de aspecten even belangrijk of hoe zou je die ranken?

Performance measures:

"A performance measure can be defined as a metric used to quantify the efficiency and/or effectiveness of an action" (Neely et al., 1995, p. 1229)

21. When is in your opinion performance evaluation useful?
22. Based on which factors would you evaluate the mail delivery performance?
23. How and when would you measure those factors?
24. Do you have any performance evaluation tools next to the official tools given by the control department?

- Wanneer is volgens jou performance evaluatie nuttig?
- Met welk factoren zou je de performance van de bezorging meten?
- Hoe en wanneer zou je deze factoren meten?
- Gebruik je naast de dashboards en het benchmarking model van controlling nog andere hulpmiddel om de bezorg performance te evalueren/ controleren

Clustering:

25. Do you think the current clustering is useful? Why?
26. What is in your opinion a perfect clustering for the mail delivery process?
27. Which aspects would you incorporate in the clustering?

- Is de huidige clustering geschikt voor de benchmarking? Waarom?
- Wat is de perfecte clustering voor het bezorgen volgens jou?
- Welke aspecten zou je voor de clustering gebruiken?

Current benchmarking model

28. How do you evaluate the current benchmarking model?
    a. What are in your opinion the strength and weaknesses?
29. How do you evaluate the benchmarking sessions?
    a. What are in your opinion the strength and weaknesses?
30. To what extent do you think is the benchmarking of mail delivery performance between the process managers of delivery useful?
31. What was the most valuable takeaway/learning of benchmarking until now? Why?
32. How would you improve it?

- Hoe beoordeel je het huidige benchmarking model?
    - Wat zijn volgens jou de sterkte en zwakte punten?
- Hoe beoordeel je de benchmarking sessies?
    - Wat zijn volgens jou de sterkte en zwakte punten?
- Hoe waardevol vind je de benchmarking?
- Wat was het meest waardevol wat je door het benchmarking model of/en sessies hebt geleerd?
- Hoe zou je het willen verbeteren?

Benchmarking model

33. What do you think are important elements of a benchmarking model?
34. How would you gain the motivation of the process managers?
35. In what frequency and form should the benchmarking be hold?

- Wat zijn belangrijke onderdelen voor een benchmarking model?
- Hoe zou je de proces manager beter kunnen motiveren om met het model te werken?
- Hoe vaak en op wat voor een manier zou je de benchmarking sessies houden?

Closure of the interview

36. Do you have any final comments or thoughts on this matter you would like to share?
    Heb je nog andere opmerkingen of gedachten die je zou willen delen?

Appendix V Cluster Attributes based on the Stakeholder Interviews

In the following we summarised all clustering attributes named by the stakeholders. The frequency that it is named is stated within the brackets. If no frequency is stated, it was named once in all the interviews.

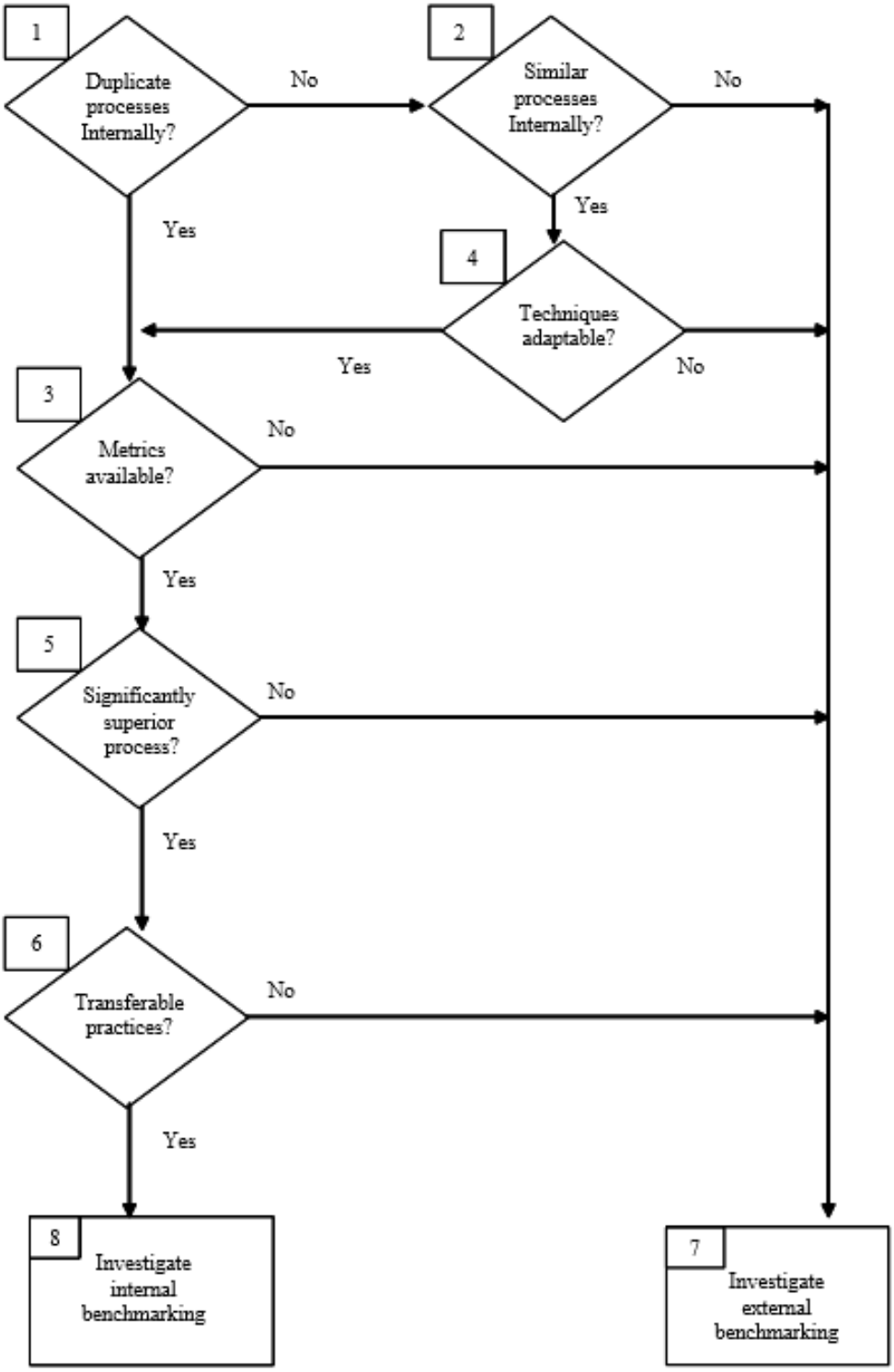
- access to mailbox (5x)
    - location of mailbox (floor level)
- interdrop (5x)
- layout of houses (5x)
    - new buildings, family houses, front yards
    - new building, shopping mall, industry, polder,
    - stairs, garden, flat building
- address arrangement (4x)
    - old cities have red or back numbers, affix
    - structure of the street: house (sub) number, one sequence or messy (sub-numbers: more risk of doing wrong, in one coherent sequence …)
- # delivery points (3x)
- customer (3x)
    - kind of customers (people who complain quickly)
        - less in rural areas, less in the north
        - rural area: people are more easy going
    - wishes of customer
    - home or companies: companies have more mail*

- Kind of mail(3x)
    - Ring packages require more time than normal mail

- volume (amount of mail) (2x)
    - time to control address
    - # mail items: occasionally there are peaks, e.g. postcode lottery
- chances that someone is home (higher in villages than in cities) (2x)
- traffic (2x)
    - delivery to depot: chances to wait due to traffic
    - parking spots, traffic lights
- labour market
- delivery points per km2
    - density of houses (deliver package to neighbour)
- #complaints compared to # delivery points
- injuries in one distract
- length of main-route
- run-up to a depot
- means of transportation
    - how much you can transport (# reloading)
    - if you go by bike or postboy: bike can fall, other movements (bike: more bundles, postboy, more or less for every address)
- weather: city is less extreme

Appendix VI Benchmarking Process Flowchart

Southard and Parente (2007) develop a flowchart for selecting between internal and external benchmarking (see Figure_Apx VI-1).



FIGURE_APX VI-1: BENCHMARKING PROCESS FLOWCHART (SOUTHARD & PARENTE, 2007)

Appendix VII AIC vs BIC

There are two main distinctions between AIC and BIC. The AIC criterion penalises the parameters less and thus often overestimate the number of parameters (Steinley, 2006). Another distinction between those criteria is the theoretical starting point. The BIC is based on the Bayesian approach (see Larsen and Marx (2012) for an outline of the Bayesian approach), which aims to identify among the candidate data the real model and selects the candidate model with the highest probability. In contrast, the AIC model is based on the Kullback-Leibler (K-L) distance, which assumes that the real model cannot be identified; however this approach aims to identify the best approximation of the real model. The idea is that each approximation tries to predict future data, however it comes along with a certain information loss. The Kullback-Leibler distance measures the relative information loss between the models and by that does not even require to know the real model. For a detailed outline of the theoretical background and calculation we refer to Kuha (2004).

Appendix VIII Analysis of the Critical Success Factors based on the Four Perspective Model

In the following we present the detailed analysis for each perspective of the four perspective model on operations strategy of Slack et al. (2010) to define the CSFs. First the top-down perspective, second the bottom-up perspective, third the market requirement perspective and finally the operations resources perspective.

1. Top-Down Perspective

The top-down perspective considers the expectations of the business on the operations. Therefore, we analyse the strategy defined by the logistic strategy department for the mail delivery.
The current mission of the mail delivery of PostNL is to establish a flexible delivery network and to develop additional tasks for the postmen to enable a cost efficient delivery with high quality. Their strategy is based on three main areas: process, conduct, management.

- **Process:** Enable a cost efficient delivery process by an optimal usage of the different means of transportation. Optimise the estimated delivery time for a tour by using (real-time) data. Frequently control and if necessary change the tour composition.
- **Conduct:** The conduct of employees, staff and line, has to be in line with the characteristics of a flexible delivery network. This means for postmen that they should have an open attitude to deliver different tours.
- **Management:** The management should be linked to the flexible logistic structure. Implementing an adequate information system for staff and line to facilitate a better information exchange.

According to the strategy one CSF is flexibility including the flexibility of an employee, but also the set-up of the delivery tours. Another CSF is the cost efficiency created through an adequate usage of means of transportation and a better delivery time estimation per tour. Considering the mission we would define quality and innovativeness in respect to new tasks as CSF, however it is not within the focus of the current strategy. Innovativeness is important, however the process manager is rather responsible for the implementation than the deployment of new tasks and it is not relevant for the benchmarking model. To sum up, from a top-down perspective the CSFs for the delivery process are flexibility and cost efficiency.

2. Bottom-Up Perspective

This perspective considers the day-to-day operations to determine the CSF. Therefore we compare theory and practice, by determining first the key elements suggest by the academic literature and subsequently compare it with the findings based on the interview with the employees from the line (postmen, team leaders, process managers) (see 3.1).
The academic literature agrees on five key factors for operational processes on a day-to-day level, which are quality, speed, dependability, flexibility and costs (Landeghem & Persoons, 2001; Slack et al., 2010). These factors can be applied to the mail delivery as well. Chan, Henry and Ralph (2009) presents a list with critical success factors for the postal service, which highly overlaps with the five key factors in operations. In the following, we discuss the five key factors briefly, relate them to the mail delivery process in general and more specific for PostNL.

- **Speed** means doing things fast. Within the mail deliver the speed would be measured as lead time between receiving the mail and delivering it to the destination (Chan et al., 2009). PostNL offers different lead-times: 24h, 48h and 72h. Considering our scope there is no differentiation, as after the mail reaches the depot or HUB there all mail has to be delivered on that day.

- **Dependability** means keeping the delivery promises. To satisfy the customers, post companies need to offer a reliable service (Chan et al., 2009). PostNL promises under the USO to delivery in the Netherlands at least 95% of the mail the next day.
- **Quality** means that you want to do things right. Good quality is defined by PostNL to delivery on time, to the correct address without damage. Therefore, the quality incorporates the performance factors speed and dependability.
- **Flexibility** means the ability to change what you are doing. Chan et al. (2009) point out that post companies respond effectively to changes if the labour can perform various tasks. As mentioned in section 2, from a more strategic level it concerns the highly decreasing volume within the mail sector. However, on a day-to-day level they have to handle the varying volume level per day and per week, but also to manage peak periods like the Christmas season. PostNL calculates one week in advance the expected volume (amount of mail per kind of mail) per day in order to enable process managers and team leaders to manage and schedule their postmen efficiently.
- **Cost** means producing cheaply in order to offer a reasonable price for the market and still gaining reasonable profit. One should not only assess cost isolated, but also in ratio to the output; Chan et al. (2009) do not only use cost (manufacturing, item carrying, storage and distribution), but also the return on assets as a measure of efficiency in utilizing assets. PostNL focuses highly on cost in their MJ dashboard; however they do not consider the utilization of assets.

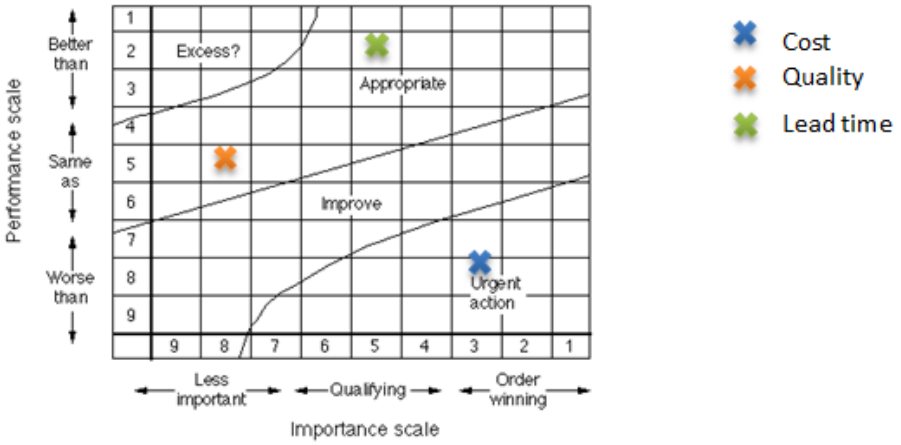
Additionally to the five key factors, Chan et al. (2009) include **innovativeness, convenience** and the **relationship** with customers, employees and partners.

- **Innovativeness** is measured by the number of new services and technology launched (Chan et al., 2009). PostNL strives for innovations; however as mentioned above not within the responsibility of the process managers.
- **Convenience** is related to offer customer postal service within close reach (Chan et al., 2009). For the main deliver network of PostNL, which is our research scope, the addresses that receive delivery are fixed and thus not relevant for the performance measurement.
- The **relationship** with customers can be measured indirectly by the percentage of change in registered customer complaints (Chan et al., 2009). Currently, PostNL measures it within the MJ dashboard based on the number and sort of complaints. The relationship with the employees is also measured indirectly at PostNL by the indicator employee commitment. The mail delivery process as defined in our scope has no external partners and thus measuring the relationship with partners is not applicable.

Overall, we can find six CSF in the academic literature which can be applied on a day-to-day level at mail delivery process: speed, dependability, quality, flexibility, cost and relationship. All of them are also emphasized by the line employees, except innovativeness (see stakeholder analysis in Section 3.1). In contrast to the literature, our findings of the stakeholder analysis show that CSF are interrelated and not always equally important; The highest priority set by the line employees is the customer satisfaction measured in customer complaints. The second priority is the quality with the focus on delivering on time. Therefore, it is critical to make a good schedule based on the volume prediction, but also to ensure flexibility among the postmen in case of unexpected variations. The third priority is the employee commitment, as motivated employees are more flexible, deliver more efficient and show a better behaviour towards customer. Finally, cost is less emphasized by the line employees as cost is only relevant if it exceeds the budget.
To sum up, based on the bottom-up perspective we can define (from high to low priority) the following CSF: customer satisfaction, quality, flexibility, employee commitment and cost. Furthermore, we can see that the CSFs can be interrelated which should be considered as well in determining the performance measures.

3. Market Requirements Perspective

This perspectives considers “what the market position requires operation to do” (Slack et al., 2010, p. 65). The market position depends on the customer needs and the competitors’ performance. In this section we sorely consider the customer needs that can be directly influenced by the performance of the operation. Thus, extra services like track and trace that cannot be improved by improving the operational performance are neglected.
Slack et al. (2010) identifies the importance of performance objectives based on the need and preference of the customers but also based on the competitive position. For assessing the need of the customer, they apply the concept of Hill (1993), who defines order winners and qualifiers to translate the customer need in operational requirements; Order qualifier are prerequisite to enter the market, they do not help to win extra business, but lose business if it falls below the standard. Order winner are used to differentiate from competitors, succeeding in them helps to win extra business. At the latter, we can see that the customer expectations also depend on what the market is currently offering. Therefore one should also evaluate how the company is performing compared to the competitors. The importance-performance matrix shows the importance of performance objective based on the importance for customers and the performance against competitors and helps us to determine the CSF based on the market requirement perspective (see Figure_Apx VIII-1).



FIGURE_APX VIII-2: PRIORITY ZONES IN THE IMPORTANCE-PERFORMANCE MATRIX (SLACK ET AL., 2010)

Looking at the mail market, we can say that customers anticipate sufficient quality of the postal companies; to get the mail to any address in the Netherlands delivered on time without any damage. Thus quality is an order qualifier, however mostly taken for granted and therefore less important. Other aspects that the customer considers are the cost for sending an item and the overall lead time. In general there are two different customer groups: Those who priorities the delivery time and choose the mail service with the fastest delivery (delivery time is then the order winner); against those who do not mind the delivery time but focuses on the price (cost is then the order winner). As an example, the business model of the Swiss Post offers two different post stamps (A, B); Stamp A costs 1 CHF and includes a delivery within 24h. Stamp B is with 0.85 CHF cheaper but includes a delivery time of a couple of days. The majority of sold stamps is with 60% stamp B.
In the Netherlands private customer cannot choose yet between cost and delivery time, but automatically get the 24h delivery based on the Universal Service Obligations. However, business customers, who make 96% of the total mail volume of PostNL, can select between a fast and more expensive or a longer and cheaper delivery. According to the research on the post marked conducted by the authority of consumers & market (ACM) in 2015, 31% of the total business mail volume is delivered within 24h. Further, the business mail volume is declining faster in the 24h delivery with around 41% between 2014 and 2015 than in the 48h or 72h delivery with around 16% between 2014 and 2015 (ACM, 2016). Thus we can say that cost is rather an order winner for the customers than the delivery time. Therefore we rank cost a little bit higher than the delivery time.

Evaluating the market position compared to the competitors, we can say that there only one main competitor on the national level for PostNL, the other competitors are small local post companies who often depend to some degree on PostNL. Therefore, we evaluate the performance of PostNL in the following on the main competitor Sandd, who has a market share of 30-35% in the business mail sector (ACM, 2016).
Judging the quality performance against Sandd, PostNL offers a slightly better performance than their main competitors in the national mail sector Sandd. Sandd delivers 96% of the mail on time and without damage to the right address (Sandd, 2016b), while PostNL is slightly more reliable with 96.4 % (PostNL, 2016a).
Looking at the lead time, PostNL offers a better service than Sandd. Sandd delivers only two days a week and can offer therefore only a 72hour mail delivery service (Sandd, 2016a). In contrast, PostNL delivery 5 days a week and offers a 24h, 48h and 72h delivery service. (ACM, 2016)
Considering the last objective cost, we can say that PostNL performs worse than their competitor Sandd. Sandd offers on average a lower price than PostNL. As mentioned above, customers select rather on price than on delivery time. Consequently, PostNL is losing some customers to Sandd, which can be seen based on the increasing market share of Sandd. While they had a market share of 25-30% in 2014, it increased to 30-35% in 2015 (ACM, 2016).
To sum up, the market requirement perspective we can derive three CSF for the operation, cost, quality and lead time, which are relevant for the market position. By evaluating those CSF on the performance against competitors and on the importance for customers, we can see that the quality and lead time are currently appropriate, but that the cost requires urgent actions. Thus cost has the highest priority followed by quality and lead time.

4. Operations Resources Perspective

This perspective considers the capabilities of the operational resources and is based on the resource-based view (RBV). RBV says that a firm can create a sustainable competitive advantage based on the core competences of their resources. To achieve this, the resources should be value creating, rare and inimitability. (Johnson et al., 2011)
However, firm's processes and activities might also be limited by their resources and capabilities. In other words, the effectiveness of the business activities and processes depends critically on the resources and capabilities. Even though, the endowment of resources on a short to medium term might be limited, managers may be able to redesign some activities or processes to exploit their current resources and capabilities more efficiently and effectively. (Ray, Barney, & Muhanna, 2004)
Thus, to assess the performance of the mail delivery process, it is important to clarify their main resources with the capabilities but also their constraints. Based on the observations and the process analysis (see Chapter 2), we can identify main resources: the hard resources as depots/HUBs and means of transportations as well as soft resources like the employees. None of those resources is rare or inimitable; however it is still critical to ensure that they are exploited as efficient as possible to create the highest value. In the following the shortly explain the resources and determine the factors that need to be considered for their exploitation.
Depots can have different sizes and thus different capacities for storing mail. Thus when deciding on the depot size, one have to estimate the numbers of houses they want to delivery from that depot. Depending on the depot location, the time for delivery can vary. This includes the time to transport mail to the depot and from the depot to the delivery point. In order to exploit the resource efficiently of depots, we have to determine an optimum trade-off between the cost of the depot (renting costs) and the cost related to the delivery time to and from the depot.
The means of transportation determines the travel speed, the volume that can be transported as well as the distance that the postman can travel. For instance riding 80km by car per day is doable; however by bike it would be not acceptable for the postman. Furthermore, there are different costs connected to each means of transportation including the acquisition as well as maintenance cost. Mostly PostNL first determines the route of a tour and subsequently the means of transportation.

However, PostNL strives for more flexible tours, thus this requires to frequently revising the means of transportation in order to ensure an efficient exploitation of that resource.
For the mail delivery process the most critical resource are the people, without the postmen PostNL would not be able to offer their service. However, this is also one of the hardest resources to manage. People are less predictable as machines, as the reliability and capability varies per postman, and the management has to ensure their satisfaction in order to keep them. With 26500 postmen, it is hard to determine the capability of each one. PostNL developed some time estimations for each delivery tours, but still allows the postman to correct it if he/she thinks it is necessary as PostNL. Thus, PostNL relies hardly on the trustworthiness and perception of the postmen. In order to exploit the resource "postmen" efficiently, it requires good management to ensure employee commitment and satisfaction, but also some sort of control mechanism which makes the time planning more objective.
To sum up, from an operational resource perspective we can see that the critical success factor is the efficient use of depots and means of transportation as well as postmen efficiency. Next to those CSF we saw, that PostNL highly depends on the perception of the postmen. In order to create more independency, a control system that enables a more objective comparison for the delivery time would be necessary.

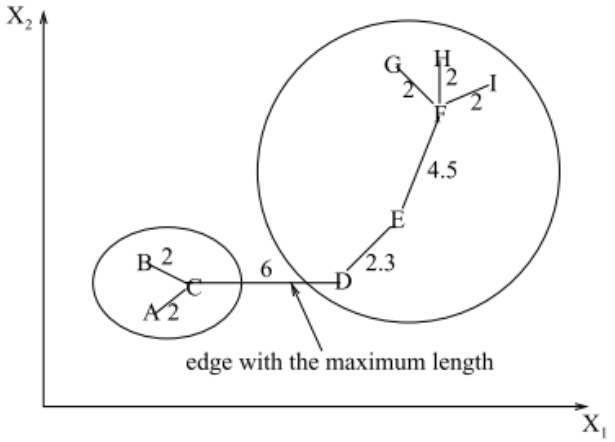
Appendix IX Clustering techniques

In this section we briefly explain graph-based and density-based clustering.

Graph-based

Prototype-based clustering is mostly expressed in algebraic constructs like the square-mean-error, however one can also base the clustering on graph theory. In graph-based clustering objectives are represented by a node in a weighted graph. Each axis of the graph represents one attribute of the object. The objects are positioned in the graph based on their values. The distance between the nodes indicates their similarity. Clusters are defined based on their connectedness and completeness. (Jain & Dubes, 1988; Tan et al., 2005b) Most commonly the minimal spanning tree (MST), introduced by Zahn (1971), is used to build clusters (see Figure_Apx IX-1). Initially all objects are connected to each other in such a way that the total weight of edge is minimised without forming a circle within that spanning tree. Subsequently, the longest edges are erased in order to form clusters. (Jain, 2010; Xu & Wunsch, 2005)

Overall graph-based clustering provides a visualisation of the data structure and enables us to already see some cluster tendencies. However, visualising more than three dimensions with a graph-based theory is difficult and thus graph-based clustering is only favourable for less than three dimensional data.



FIGURE_APX IX-1: MINIMAL SPANNING TREE APPLIED TO NINE POINTS (Jain, Murty, et al., 2000, p. 17)

Density-based

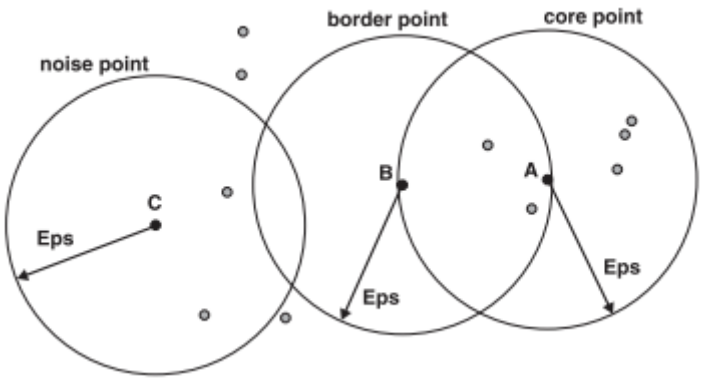
In density-based clustering, clusters are defined as a dense region of objects separated by low density regions. This provides three main advantages; firstly, density-based clustering is highly flexible in their cluster shape as clusters can grow in any direction with sufficient density. Thus, it is favourable for intertwined or irregular data structures. Secondly, density-based clustering is not affected by noise and outliers as those mostly do not form a dense area and thus are neglected in the clustering. Thirdly, it can be applied to data points as well as extended objects. (Berkhin, 2006; Tan et al., 2005b)

However, in certain circumstances those properties can also be less favourable; for instance density-based clustering would not detect two clusters with significant different, but still above the user-specified threshold, densities, if they are adjacent. Furthermore, as it requires a metric space it is most efficient with low-dimensional data with numerical attributes, also called spatial data. High dimensional data is less favourable as it is harder to define density. Finally, if the density of cluster is highly varying, objects or clusters might be neglected, as it is harder for the density-based approach to detect the clusters between the noises. (Berkhin, 2006)

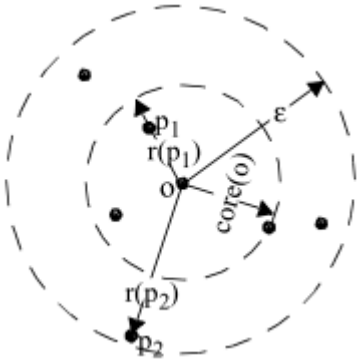
Similar to the notion of similarity, there are also many notions for density. One simple and effective density-based clustering algorithm is DBSCAN (density-based spatial clustering of applications with noise) which uses the centre-based approach for measuring density; density is measured from a

particular point and based on the number of points within a certain radius (see Figure_Apx IX-2). The algorithm differentiates between three kinds of points. 1) a point is called a core point when it has within a certain radius, Eps., more than MinPts points, where MinPts is the user specified threshold of neighbouring points. 2) if a point is not a core point, but within the radius Esp. of a core point, it is called a boarder point. A border point can be within the radius of multiple core points. 3) If a point is neither a border nor a core point, it is a so called noise point. (Tan et al., 2005b)
Based on this definitions we can follow the algorithm of DBSCAN (Tan et al., 2005b, p. 528):
1: Label all points as core, border, or noise points.
2: Eliminate noise points
3: Put an edge between all core points that are within Eps of each other.
4: Make each group of connected core points into a separate cluster.
5: Assign each border point to one of the clusters of its associated core points.

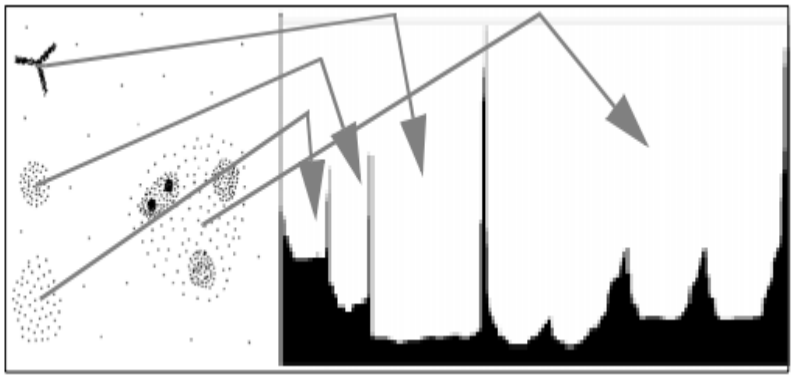
Looking at the steps, we can see that the outcome is highly sensitive to the user-defined values for the radius Esp. and the number of points MinPts. Furthermore, in order to perform step 5 one has to define how to handle ties, which is the case if a border point is within the radius of multiple core points of different clusters. Overall, the computational time depends on the number of points and the search within an Eps-neighbourhood which results in the worst case in O (m^2).
A less sensitive algorithm, which can cope with different densities is the Ordering Points To identify the Clustering Structure algorithm (OPTICS) presented by Ankerst et al. (1999) (Jain & Dubes, 1988). It uses similar to the DBSCAN information on core points and border points based on the radius Esp. and the number of neighbours MinPts. In contrast to DBSCAN, it displays for each border point its exact distance to the core point (reachability distance) and also displays for each core point the core-distance, which is the minimum radius required for that point to reach the MinPts neighbours. Instead of a clustering it plots the distances of each object in a reachability graph (see Figure_Apx IX-4). By that it visualise the density allocation of objects within the data space. Thus, overall OPTICS can be good tool for assessing the cluster structure, however it does not provide a final clustering. For an exact outline of the algorithm see (Ankerst et al., 1999). Density based clustering is quite risky when basing the analysis on a sample, where the densities might not be representative.



FIGURE_APX IX-2: CENTRE-BASED DENSITY TECHNIQUE. IN THIS CASE A IS A CORE POINT, B A BORDER POINT AND C A NOISE POINT (Tan et al., 2005b, p. 528)



FIGURE_APX IX-3: CORE- AND REACHABILITY DISTANCE (R) (ANKERST ET AL., 1999, P. 4)



FIGURE_APX IX-4: REACHABILITY-PLOT (Ankerst et al., 1999, p. 6)

Appendix X Data Analysis

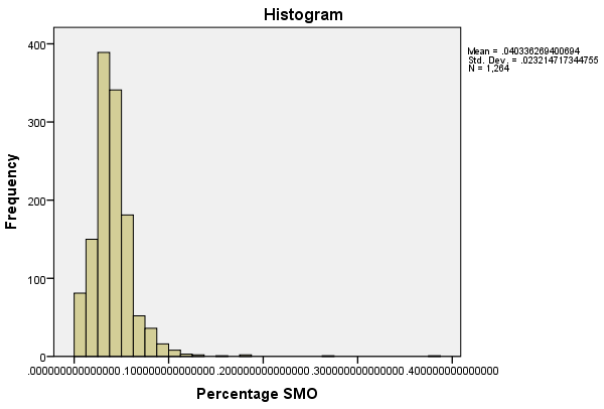
Distribution SMO Mail

During the first 9 weeks of 2017 we collected data on the volume and kind of mail for the delivery area of Utrecht. In the following we analyse the frequency of SMO mail and the ratio of SMO mail and the remaining (small and big) mail, which fits through the mail box.
In order to assess the frequency of SMO mail, we compare the number of SMO mail per delivery point among all PC5 areas within the delivery area Utrecht. The majority of delivery points (98%) get one SMO mail item within a week (see Table_Apx X-1). Considering that the mail is delivered five times a week, it supports our assumption that the frequency is in general quite low.

Bin	*Frequency*	Percentage
0	18	1.5%
1	1152	97.7%
2	6	0.5%
More	3	0.3%
Total	1179	100%

TABLE_APX X-1: SMO MAIL PER DELIVERY POINT PER WEEK

To compare if the ratio between the remaining mail and SMO mail is the same between the different PC5 areas or if there are areas that tend to have more, we evaluate the percentage of SMO mail with the PC5 areas. Given the histogram in Figure_Apx X-1 , we can see that the percentage of SMO is not equal between the PC5 areas. We can see a right tail, indicating that there are PC5 areas with extreme high SMO proportions. Given that the SMO mail can have a high impact on the delivery time and the distribution with a right tail, it is suitable to incorporate the kind of mail within the performance measurement.



FIGURE_APX X-2: HISTOGRAM OF THE PERCENTAGE OF SMO MAIL

Bin	*Frequency*	Percentage
0.05	153	12%
0.1	626	50%
0.2	448	35%
0.3	31	2%
More	6	0%
Total	1264	1

TABLE_APX X-2: PERCENTAGE OF SMO MAIL FOR PC5 AREAS

Normality Assessment

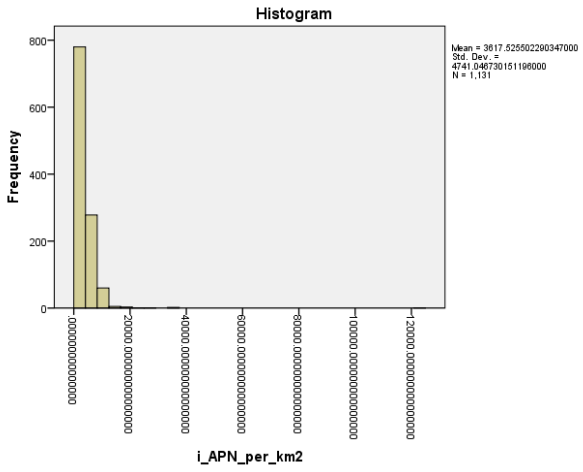
In order to assess the normality of the attributes, we conduct a normality test with SPSS and also evaluated the histogram and boxplot of each attribute given each scenario (I =infrastructure, o= off-peak, p= peak),

Tests of Normality

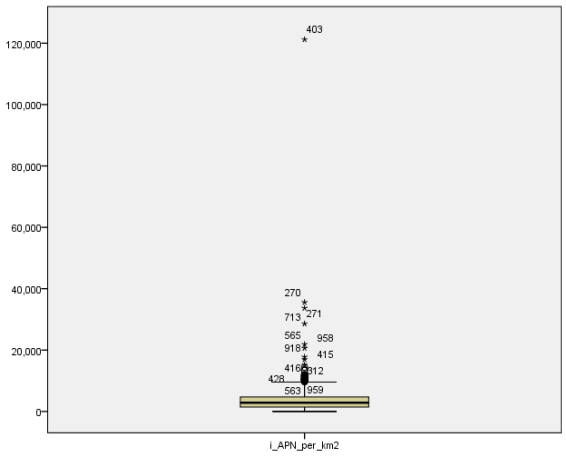
	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
i_APN_per_km2	.223	1131	.000	.449	1131	.000
i_Interdrop_4	.272	1131	.000	.453	1131	.000
i_APN/m	.337	1131	.000	.169	1131	.000
Minor-route	.090	1131	.000	.944	1131	.000
Total_Length	.144	1131	.000	.760	1131	.000
p_APN/km2	.153	1131	.000	.637	1131	.000
p_Interdrop	.344	1131	.000	.207	1131	.000
p_APN/m	.279	1131	.000	.250	1131	.000
o_APN/km2	.136	1131	.000	.688	1131	.000
o_interdrop	.371	1131	.000	.149	1131	.000
o_APN/m	.269	1131	.000	.273	1131	.000

a. Lilliefors Significance Correction

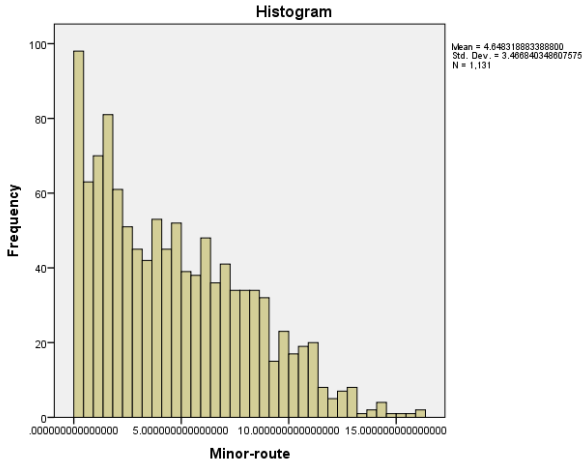
TABLE_APX X-3: : NORMAILTY TEST SPSS



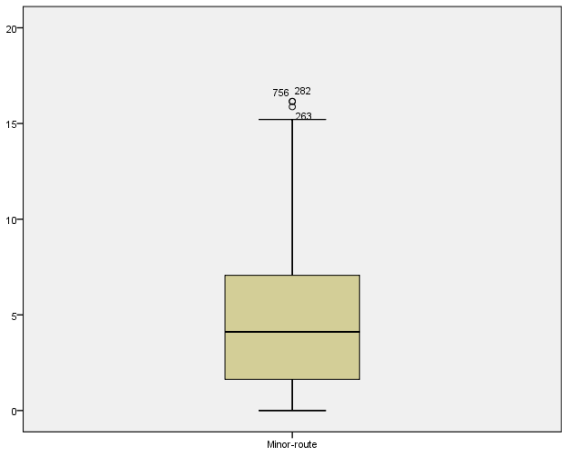
FIGURE_APX X-2: HISTOGRAM APN/KM² (INFRASTRUCTURE SCENARIO)



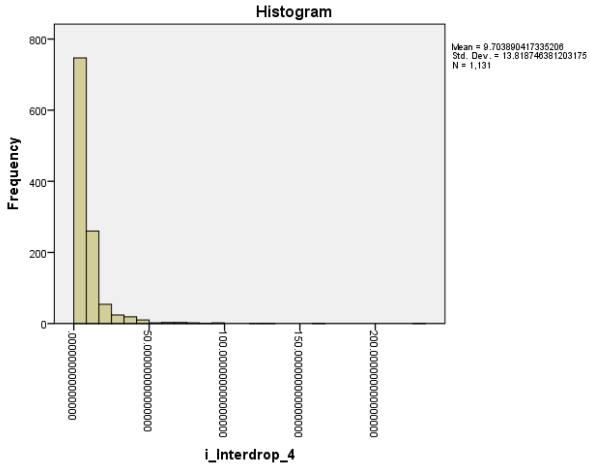
FIGURE_APX X-3: BOXPLOT APN/KM² (INFRASTRUCTURE SCENARIO)



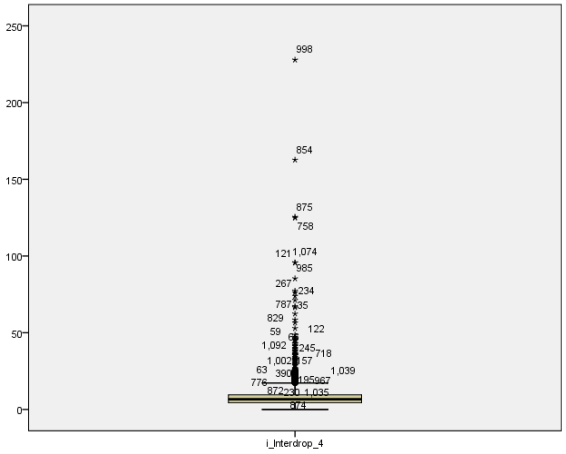
FIGURE_APX X-4: HISTOGRAM MINOR-ROUTE



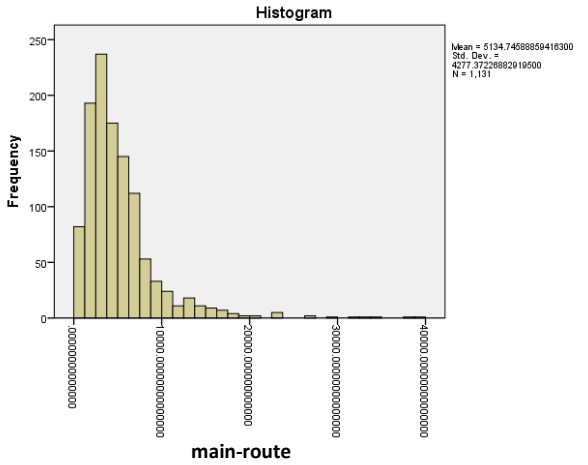
FIGURE_APX X-5: BOXPLOT MINOR-ROUTE



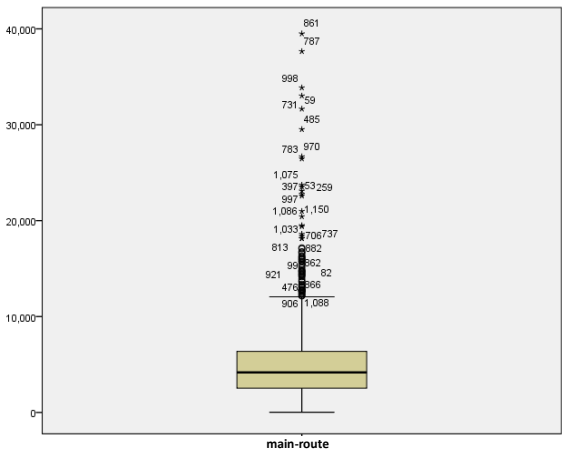
FIGURE_APX X-6: HISTOGRAM INTERDROP (INFRASTRUCTURE SCENARIO)



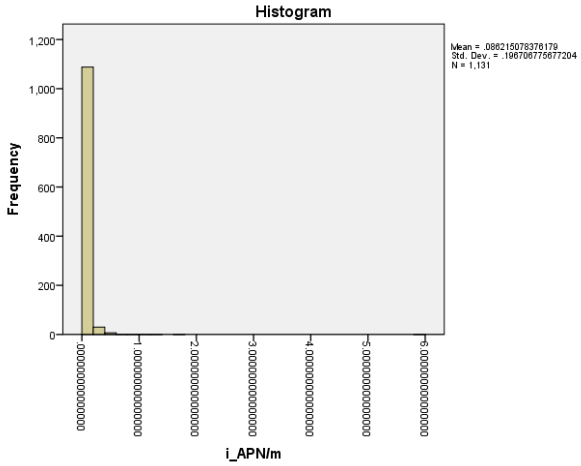
FIGURE_APX X-7: BOXPLOT INTERDROP (INFRASTRUCTURE SCENARIO)



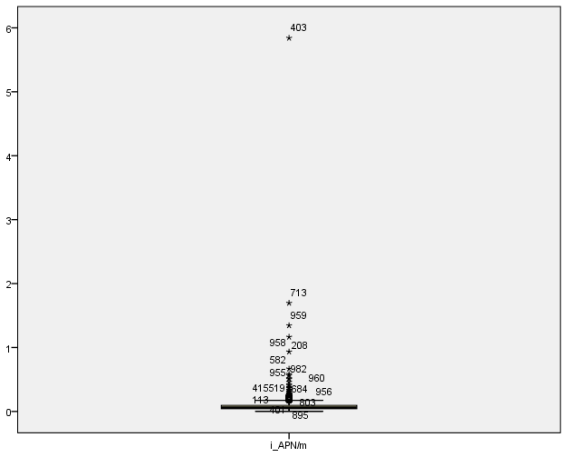
FIGURE_APX X-8: HISTOGRAM MAIN-ROUTE



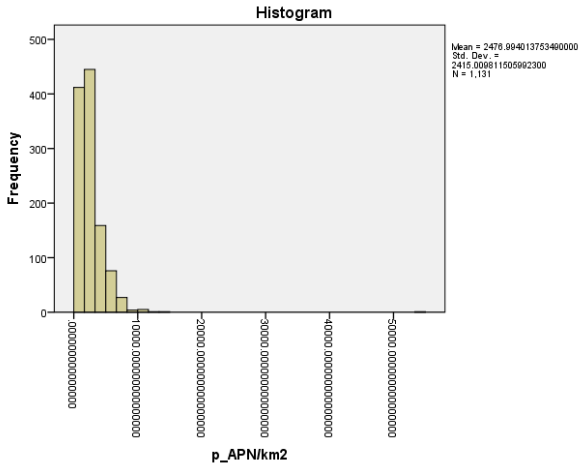
FIGURE_APX X-9: BOXPLOT MAIN-ROUTE



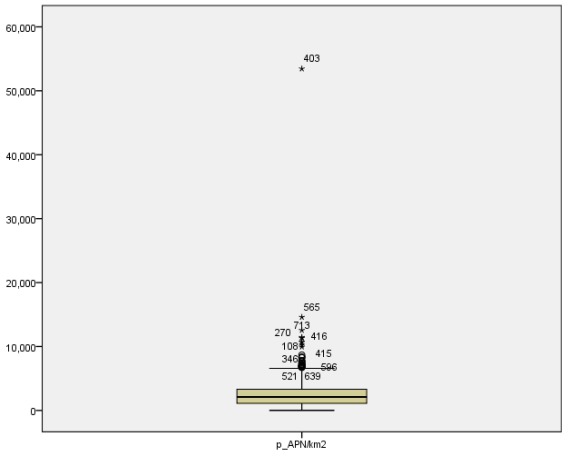
FIGURE_APX X-4: HISTOGRAM APN/M (INFRASTRUCTURE SCENARIO)



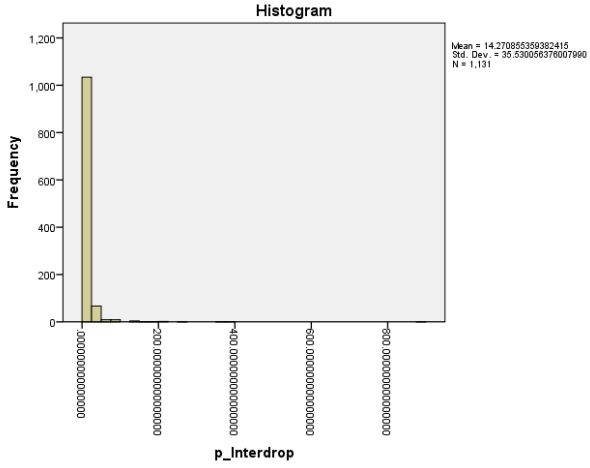
FIGURE_APX X-3: BOXPLOT APN/M (INFRASTRUCTURE SCENARIO)



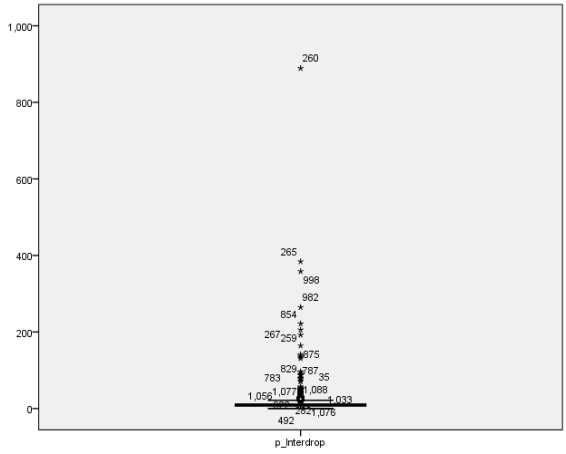
FIGURE_APX X-6: HISTOGRAM APN/KM² (PEAK SCENARIO)



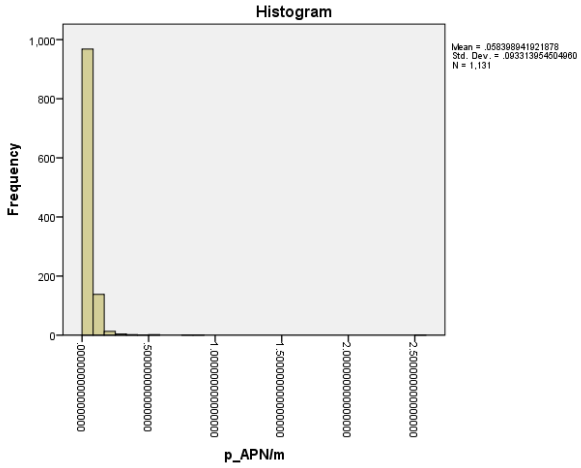
FIGURE_APX X-5: BOXPLOT APN/KM² (PEAK SCENARIO)



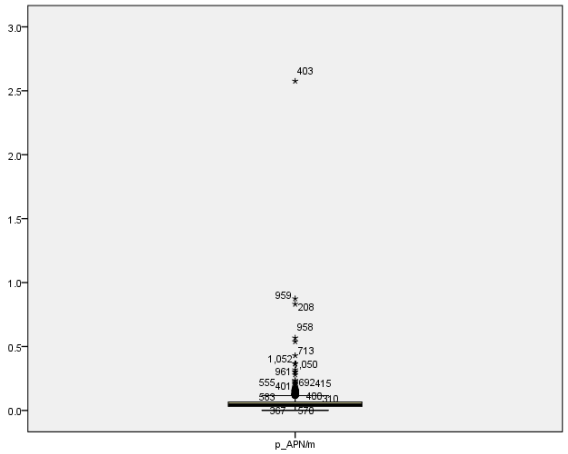
FIGURE_APX X-8: INTERDROP (PEAK SCENARIO)



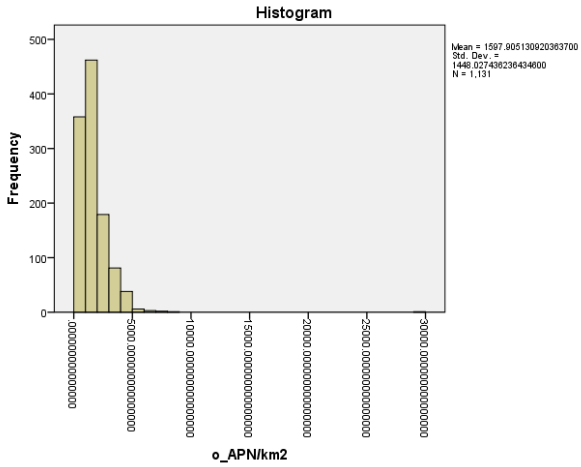
FIGURE_APX X-7: BOXPLOT INTERDROP (PEAK SCENARIO)



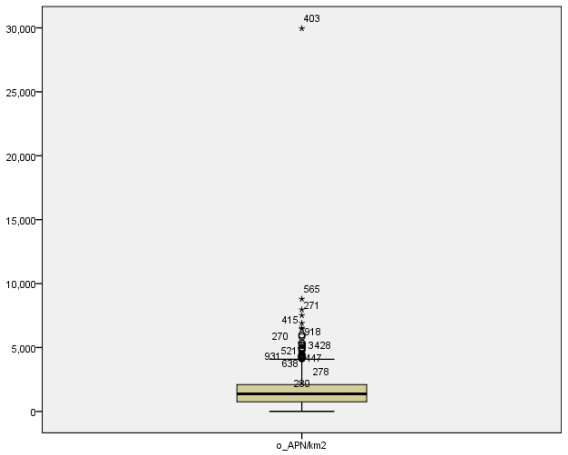
FIGURE_APX X-10: HISTOGRAM APN/M (PEAK SCENARIO)



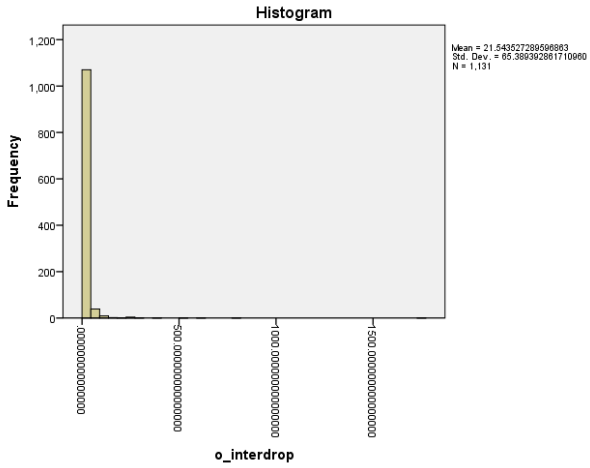
FIGURE_APX X-9: BOXPLOT APN/M (PEAK SCENARIO)



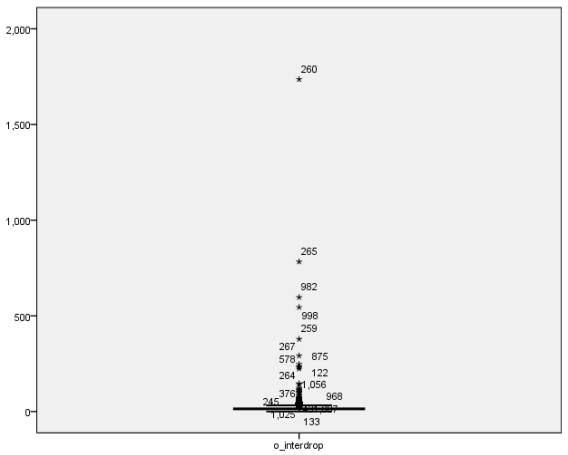
FIGURE_APX X-12: HISTOGRAM APN/KM² (OFF-PEAK SCENARIO)



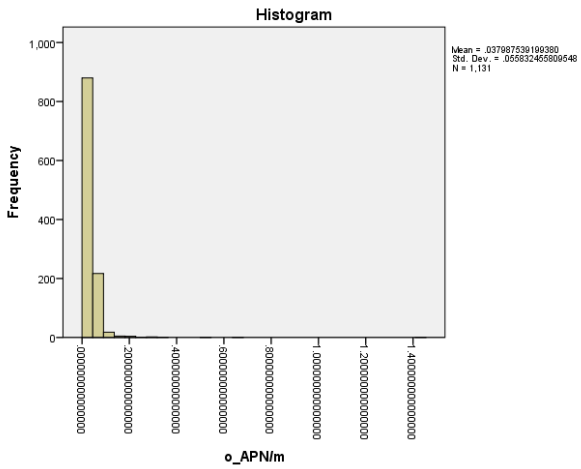
FIGURE_APX X-11: BOXPLOT APN/KM² (OFF-PEAK SCENARIO)



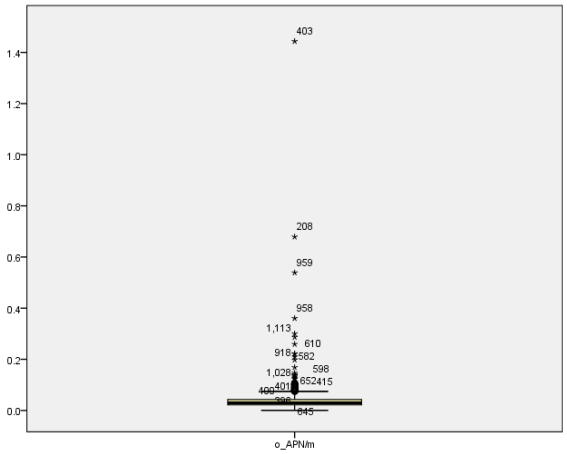
FIGURE_APX X-14: HISTOGRAM INTERDROP (OFF-PEAK SCENARIO)



FIGURE_APX X-13: BOXPLOT INTERDROP (OFF-PEAK SCENARIO)



FIGURE_APX X-16: HISTOGRAM APN/M (OFF-PEAK SCENARIO)



FIGURE_APX X-15: BOXPLOT APN/M (OFF-PEAK SCENARIO)

Linear correlation between potential Cluster Attributes

To assess the relation between the potential attributes we measure the pairwise linear correlation between them given the Pearson correlation coefficient (see Table_Apx X-).

TABLE_APX X-4: CORRELATION BETWEEN ATTRIBUTES GIVEN SCENARIO I, P, O

Correlations

		Minor-route	Main-route	i_APN_per_km2	i_Interdrop_4	i_APN/m	p_APN/km2	p_Interdrop	p_APN/m	o_APN/km2	o_interdrop	o_APN/m
Minor-route	Pearson Correlation	1	0.055	-.185**	0.019	-.109**	-.211**	-0.012	-.126**	-.216**	-0.014	-.128**
	Sig. (2-tailed)		0.060	0.000	0.523	0.000	0.000	0.670	0.000	0.000	0.627	0.000
	N	1165	1165	1165	1165	1165	1165	1165	1165	1165	1165	1165
Main-route	Pearson Correlation	0.055	1	-.283**	.173**	-.193**	-.344**	.062*	-.245**	-.357**	0.046	-.256**
	Sig. (2-tailed)	0.060		0.000	0.000	0.000	0.000	0.035	0.000	0.000	0.114	0.000
	N	1165	1165	1165	1165	1165	1165	1165	1165	1165	1165	1165
i_APN_per_km2	Pearson Correlation	-.185**	-.283**	1	-.184**	.818**	.966**	-.088**	.788**	.946**	-.071*	.744**
	Sig. (2-tailed)	0.000	0.000		0.000	0.000	0.000	0.003	0.000	0.000	0.015	0.000
	N	1165	1165	1165	1165	1165	1165	1165	1165	1165	1165	1165
i_Interdrop_4	Pearson Correlation	0.019	.173**	-.184**	1	-.096**	-.237**	.665**	-.129**	-.251**	.585**	-.133**
	Sig. (2-tailed)	0.523	0.000	0.000		0.001	0.000	0.000	0.000	0.000	0.000	0.000
	N	1165	1165	1165	1165	1165	1165	1165	1165	1165	1165	1165
i_APN/m	Pearson Correlation	-.109**	-.193**	.818**	-.096**	1	.727**	-0.042	.976**	.688**	-0.032	.948**
	Sig. (2-tailed)	0.000	0.000	0.000	0.001		0.000	0.154	0.000	0.000	0.268	0.000
	N	1165	1165	1165	1165	1165	1165	1165	1165	1165	1165	1165
p_APN/km2	Pearson Correlation	-.211**	-.344**	.966**	-.237**	.727**	1	-.116**	.728**	.996**	-.095**	.694**
	Sig. (2-tailed)	0.000	0.000	0.000	0.000	0.000		0.000	0.000	0.000	0.001	0.000
	N	1165	1165	1165	1165	1165	1165	1165	1165	1165	1165	1165
p_Interdrop	Pearson Correlation	-0.012	.062*	-.088**	.665**	-0.042	-.116**	1	-.065*	-.123**	.995**	-.069*
	Sig. (2-tailed)	0.670	0.035	0.003	0.000	0.154	0.000		0.027	0.000	0.000	0.019
	N	1165	1165	1165	1165	1165	1165	1165	1165	1165	1165	1165
p_APN/m	Pearson Correlation	-.126**	-.245**	.788**	-.129**	.976**	.728**	-.065*	1	.697**	-0.054	.992**
	Sig. (2-tailed)	0.000	0.000	0.000	0.000	0.000	0.000	0.027		0.000	0.066	0.000
	N	1165	1165	1165	1165	1165	1165	1165	1165	1165	1165	1165
o_APN/km2	Pearson Correlation	-.216**	-.357**	.946**	-.251**	.688**	.996**	-.123**	.697**	1	-.101**	.668**
	Sig. (2-tailed)	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000		0.001	0.000
	N	1165	1165	1165	1165	1165	1165	1165	1165	1165	1165	1165
o_interdrop	Pearson Correlation	-0.014	0.046	-.071*	.585**	-0.032	-.095**	.995**	-0.054	-.101**	1	-.058*
	Sig. (2-tailed)	0.627	0.114	0.015	0.000	0.268	0.001	0.000	0.066	0.001		0.049
	N	1165	1165	1165	1165	1165	1165	1165	1165	1165	1165	1165
o_APN/m	Pearson	-.128**	-.256**	.744**	-.133**	.948**	.694**	-.069*	.992**	.668**	-.058*	1
	Sig. (2-	0.000	0.000	0.000	0.000	0.000	0.000	0.019	0.000	0.000	0.049	
	N	1165	1165	1165	1165	1165	1165	1165	1165	1165	1165	1165

**. Correlation is significant at the 0.01 level (2-tailed).

*. Correlation is significant at the 0.05 level (2-tailed).

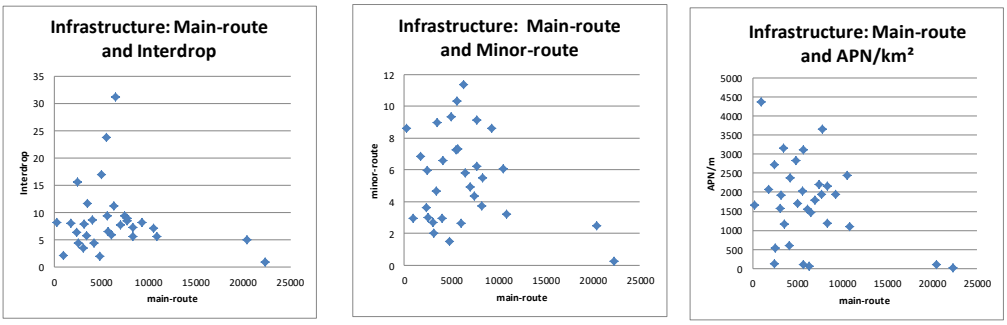
relevant for Scenario p (peak day)

relevant for Scenario o (off-peak day)

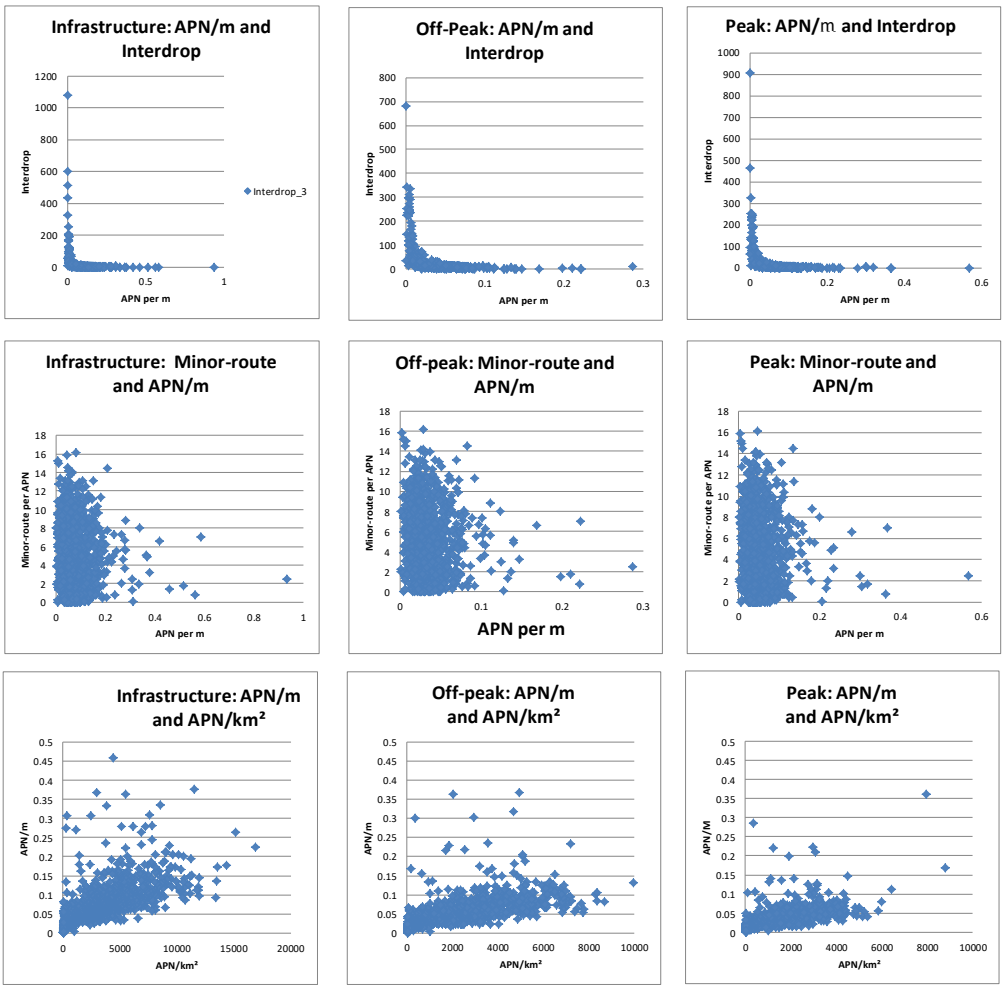
relevant for Scenario i (infrastructure)

Pairwise relation with APN/m

Looking at Figure_Apx X-17 we can conclude that there is no clear relation between the main-route per PC5 area and the remaining attributes interdrop, minor-route and APN/km². Therefore, we set, similar to APN/km², the main-route in relation to the number of APN (APN/m). Considering Figure_Apx X-19, we can see that the patterns between APN/m and interdrop as well as minor-route are similar to the one of APN/km². The high similarity can be explained by the high positive correlation coefficient between APN/m and APN/km², which is 0.818 (infrastructure scenario), 0.728 (peak scenario) and 0.668 (off-peak scenario) (see Table_Apx X-4).



FIGURE_APX X-18: RELATION OF MINOR-ROUTE, INTERDROP AND APN/KM² WITH MAIN-ROUTE



FIGURE_APX X-19: RELATION BETWEEN APN/M AND INTERDROP, MINOR-ROUTE AND APN/KM²

Test Linear Regression Peak and Off-Peak

Based on the Pearson correlation coefficient, we saw a high correlation between peak and off-peak day for each attribute. To determine their linear relationship, we conduct a linear regression of interdrop, APN/m and APN/km² which all depend on the hit-chance.

For the hit-chance we have with p=0 a significant linear relationship between off-peak and peak. Given the R^2, we can say that 88.7% of the variance of a peak day can be explained by the hit-chance of an off-peak day. Thus, considering that we use for peak and off-peak day the same constant factor of interdrop, APN/m and APN/km² and multiply it with the representative hit-chance to gain the estimator for peak and off-peak day (see Table 5.7), we expect a high and significant linear relationship as well. With a p-value of zero and a R^2 higher than 98% for of interdrop, APN/m and APN/km² we can confirm this assumption (see SPSS output below).

SUMMARY OUTPUT Hit-chance

Regression Statistics	
Multiple R	0.941893851
R Square	0.887164026
Adjusted R Square	0.887064786
Standard Error	0.03246111
Observations	1139

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	9.419840024	9.41984	8939.574	0
Residual	1137	1.198083775	0.001054		
Total	1138	10.6179238			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	0.325085207	0.004544781	71.52934	0	0.316168107	0.334002307	0.316168107	0.334002307
X Variable 1	0.832833585	0.008808457	94.54932	0	0.815550929	0.85011624	0.815550929	0.85011624

SUMMARY OUTPUT Interdrop

Regression Statistics	
Multiple R	0.99229398
R Square	0.984647342
Adjusted R Square	0.984633839
Standard Error	3.136735038
Observations	1139

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	717485.7473	717485.7	72921.84	0
Residual	1137	11187.06432	9.839107		
Total	1138	728672.8116			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	-1.117319683	0.110247612	-10.1346	3.61E-23	-1.333631296	-0.901008069	-1.333631296	-0.901008069
X Variable 1	0.760531093	0.00281636	270.0404	0	0.755005246	0.766056939	0.755005246	0.766056939

SUMMARY OUTPUT APN/m

Regression Statistics	
Multiple R	0.994531095
R Square	0.989092099
Adjusted R Square	0.989082506
Standard Error	0.003859601
Observations	1139

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	1.535822256	1.535822	103099.4	0
Residual	1137	0.016937347	1.49E-05		
Total	1138	1.552759603			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	-0.002344235	0.000205849	-11.3881	1.58E-28	-0.002748122	-0.001940348	-0.002748122	-0.001940348
X Variable 1	1.598893628	0.004979567	321.0909	0	1.589123456	1.608663799	1.589123456	1.608663799

SUMMARY OUTPUT APN/km²

Regression Statistics	
Multiple R	0.996183135
R Square	0.992380839
Adjusted R Square	0.992374138
Standard Error	156.4952629
Observations	1139

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	3626887456	3.63E+09	148092	0
Residual	1137	27846002.43	24490.77		
Total	1138	3654733459			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	-62.11245627	7.806298864	-7.95671	4.25E-15	-77.42882525	-46.79608729	-77.42882525	-46.79608729
X Variable 1	1.585182784	0.004119206	384.8273	0	1.577100685	1.593264883	1.577100685	1.593264883

PC 5 with an extreme Interdrop

Given the scatterplot we can find five PC5 areas with an extreme high interdrop. We ensured that this is no data mistake by evaluating the distance between houses with the geographical map system Geodan and Google Maps. Evaluating the values of the attributes we can see that those areas are highly varying with an interdrop range of around 800m and a hit-chance range of 0.8% during peak days (see Table_Apx X-). Thus they do not form a "natural cluster"; therefore we consider them as outliers and exclude them before applying the cluster algorithms.

PC5	Interdrop	Hit-chance peak	Hit-chance off-peal
3404P	1081.85	0.837132	0.629843
3612N	604.3333733	0.769876	0.566295
3562M	511.0000311	0.126985	0.067568
3565M	433.3333571	0.749104	0.580341
3545E	325.6666832	0.770585	0.673841
3723M	253.6415208	0.994419	0.965774
3734M	247.8461659	0.971884	0.869413
3712B	227.8370968	0.635686	0.417907

TABLE_APX X-5: ATTRIBUTES OF THE AREAS WITH AN EXTREME INTERDROP (>220M)

Appendix XI Cluster Outcome

Paired Samples t-Test for x-Means Clustering

In order to test if the cluster assignment between peak and off-peak days is significant difference, we conduct a paired t-test given cluster membership during peak and off-peak days for Scenario 1 and 2. While our H0 hypothesis is that there is no significant difference between the cluster membership in peak and off-peak days, the H1 hypothesis is that there is a significant difference.
Looking at the SPSS outcome (see Figure_Apx XI-1) and taking a 95% confidence interval (α=5%), we can say that with a significance level of 0.318, we cannot reject the H0 hypothesis for Scenario 1, however with a significance level of 0 we can reject the H0 hypothesis for Scenario 2. Thus, we can assume that there is no significant difference in the cluster membership in peak and off-peak in Scenario 1, while there is a significant difference for Scenario 2. However, with a mean difference of -0.042, this difference is quite small.

Paired Samples Statistics

		Mean	N	Std. Deviation	Std. Error Mean
Pair 1	2p	2.55	1140	.926	.027
	2o	2.59	1140	.937	.028
Pair 2	1o	1.41	1140	.493	.015
	1p	1.41	1140	.493	.015

Paired Samples Correlations

		N	Correlation	Sig.
Pair 1	2p & 2o	1140	.937	.000
Pair 2	1o & 1p	1140	.998	.000

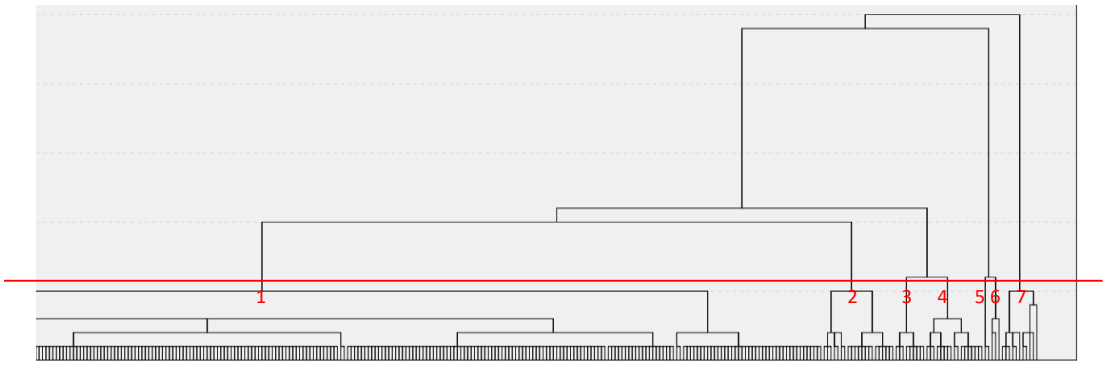
Paired Samples Test

		Paired Differences					t	df	Sig. (2-tailed)
		Mean	Std. Deviation	Std. Error Mean	95% Confidence Interval of the Difference Lower	Upper			
Pair 1	2p - 2o	-.042	.330	.010	-.061	-.023	-4.309	1139	.000
Pair 2	1o - 1p	.001	.030	.001	-.001	.003	1.000	1139	.318

FIGURE_APX XI-1: PAIRED T-TEST ON CLUSTER MEMBERSHIP GIVEN PEAK AND OFF-PEAK AND X-MEANS ALGORITHM

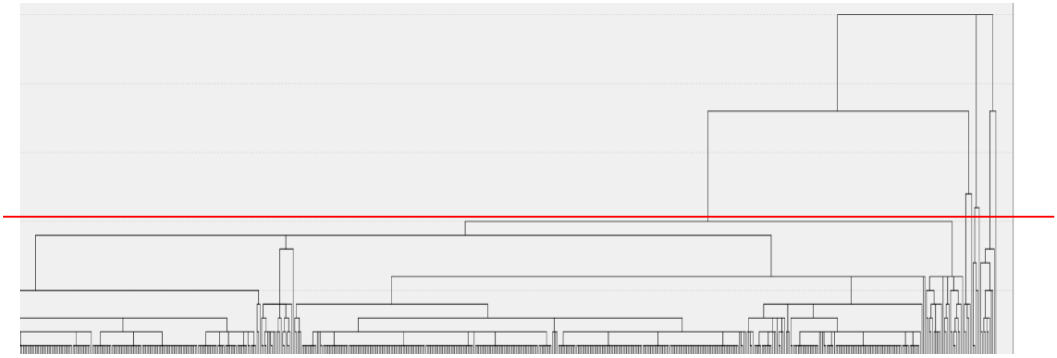
Dendrogram of the Hierarchical Clustering

Hierarchical clustering with centroid linkage is used in the TwoStep clustering. In the following figures we show the dendrogram of the last few merges. Furthermore, in Table_Apx XI-1 we provide the percentages of objects within a cluster given the red stopping line. For case 2p, we provide a detailed description on how to read it. For the remaining three, we only provide the percentage of the clusters and the last merges of the dendrogram.

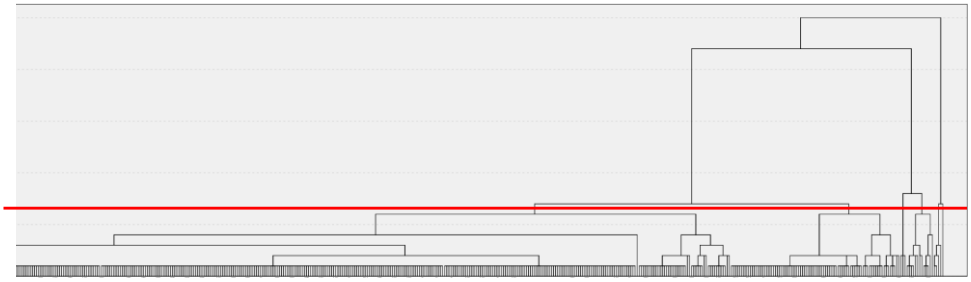


FIGURE_APX XI-2: LAST MERGES OF THE DENDROGRAM CASE 2P

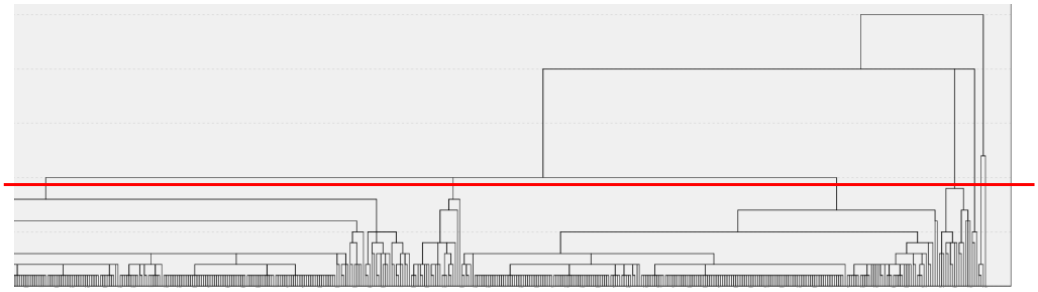
In Figure_Apx XI-2, we show partly the dendrogram of Case 2p, visualising the last merges of the clusters. Each vertical grey line indicates a cluster, while the horizontal indicates the merging of two clusters. The red line is the stopping line if we want to have 7 clusters as it goes through seven vertical lines (see red numbers). Going from cluster 1 to 7, the percentage of objects within that cluster are 94.20%, 1.76%, 0.97%, 1.58%, 0.26%, 0.35% and 0.88% respectively (see Table_Apx XI-1).



FIGURE_APX XI-3: LAST MERGES OF THE DENDOGRAM CASE 1P



FIGURE_APX XI-4: LAST MERGES OF THE DENDROGRAM CASE 2O



FIGURE_APX XI-5: LAST MERGES OF THE DENDROGRAM CASE 1O

Cluster	Case 1p	Case 2p	Case 1o	Case 2o
1	95.50%	94.20%	76.80%	91.04%
2	2.10%	1.76%	1.93%	7.03%
3	0.07%	0.97%	19.16%	0.09%
4	0.04%	1.58%	0.88%	1.23%
5	0.02%	0.26%	0.35%	0.53%
6	0.08%	0.35%	0.35%	0.09%
7	0.02%	0.88%	0.53%	-

TABLE_APX XI-1: PERCENTAGE OF OBJECTS IN CLUSTERS FOR EACH CASE

Single-Link Hierarchical Clustering

In Table_Apx XI-2 we show the cluster proportions of single-link hierarchical clustering given the number of clusters. We can see that independent of the number of clusters, there is always one dominating cluster, which includes at least 95% of the PC5 areas. In the current benchmarking the whole delivery areas are compared, however those are highly heterogeneous. Comparing 95% of one delivery area with others would not improve the problem and thus this clustering technique is unsuitable for our purpose.

TABLE_APX XI-2: CLUSTER PROPORTIONS OF THE SINGLE-LINK CLUSTERING GIVEN THE NUMBER OF CLUSTERS

Scenario	Number of Clusters	Cluster proportion (in %)
1p	12	0.982, 0.002, 0.006, 0.001, 0.002, 0.002, 0.001, 0.001, 0.001, 0.001, 0.001, 0.001
	11	0.983, 0.002, 0.006, 0.001, 0.002, 0.002, 0.001, 0.001, 0.001, 0.001, 0.001, 0
	10	0.984, 0.002, 0.006, 0.001, 0.002, 0.002, 0.001, 0.001, 0.001, 0.001, 0, 0
	9	0.986, 0.002, 0.006, 0.001, 0.002, 0.001, 0.001, 0.001, 0.001, 0, 0, 0
	8	0.987, 0.002, 0.006, 0.001, 0.002, 0.001, 0.001, 0.001, 0, 0, 0, 0
	7	0.989, 0.002, 0.006, 0.001, 0.001, 0.001, 0.001, 0, 0, 0, 0, 0
	6	0.989, 0.002, 0.006, 0.001, 0.001, 0.001, 0, 0, 0, 0, 0, 0
	5	0.989, 0.002, 0.007, 0.001, 0.001, 0, 0, 0, 0, 0, 0, 0
	4	0.996, 0.002, 0.001, 0.001, 0, 0, 0, 0, 0, 0, 0, 0
	3	0.998, 0.001, 0.001, 0, 0, 0, 0, 0, 0, 0, 0, 0
	2	0.999, 0.001, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0
1o	12	0.981, 0.002, 0.001, 0.001, 0.008, 0.001, 0.001, 0.003, 0.001, 0.001, 0.001, 0.001
	11	0.982, 0.002, 0.001, 0.008, 0.001, 0.001, 0.003, 0.001, 0.001, 0.001, 0.001, 0
	10	0.982, 0.002, 0.001, 0.008, 0.001, 0.003, 0.001, 0.001, 0.001, 0.001, 0, 0
	9	0.984, 0.001, 0.008, 0.001, 0.003, 0.001, 0.001, 0.001, 0.001, 0, 0, 0
	8	0.985, 0.001, 0.008, 0.001, 0.003, 0.001, 0.001, 0.001, 0, 0, 0, 0
	7	0.986, 0.001, 0.008, 0.001, 0.003, 0.001, 0.001, 0, 0, 0, 0, 0
	6	0.987, 0.001, 0.008, 0.001, 0.003, 0.001, 0, 0, 0, 0, 0, 0
	5	0.988, 0.008, 0.001, 0.003, 0.001, 0, 0, 0, 0, 0, 0, 0
	4	0.989, 0.008, 0.003, 0.001, 0, 0, 0, 0, 0, 0, 0, 0
	3	0.996, 0.003, 0.001, 0, 0, 0, 0, 0, 0, 0, 0, 0
	2	0.999, 0.001, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0
2p	12	0.963, 0.015, 0.004, 0.008, 0.002, 0.001, 0.002, 0.003, 0.001, 0.001, 0.001, 0.001
	11	0.963, 0.015, 0.004, 0.008, 0.004, 0.001, 0.002, 0.001, 0.001, 0.001, 0.001, 0
	10	0.963, 0.015, 0.004, 0.008, 0.005, 0.001, 0.002, 0.001, 0.001, 0.001, 0, 0
	9	0.963, 0.015, 0.004, 0.008, 0.005, 0.001, 0.002, 0.001, 0.002, 0, 0, 0
	8	0.978, 0.004, 0.008, 0.005, 0.001, 0.002, 0.001, 0.002, 0, 0, 0, 0
	7	0.978, 0.004, 0.008, 0.005, 0.003, 0.002, 0.001, 0, 0, 0, 0, 0
	6	0.978, 0.004, 0.008, 0.005, 0.003, 0.002, 0, 0, 0, 0, 0, 0
	5	0.986, 0.004, 0.005, 0.003, 0.002, 0, 0, 0, 0, 0, 0, 0
	4	0.986, 0.01, 0.003, 0.002, 0, 0, 0, 0, 0, 0, 0, 0
	3	0.996, 0.003, 0.002, 0, 0, 0, 0, 0, 0, 0, 0, 0
	2	0.998, 0.002, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0
2o	12	0.975, 0.006, 0.003, 0.007, 0.001, 0.002, 0.003, 0.001, 0.001, 0.001, 0.001, 0.001
	11	0.975, 0.006, 0.004, 0.007, 0.001, 0.002, 0.003, 0.001, 0.001, 0.001, 0.001, 0
	10	0.975, 0.007, 0.004, 0.007, 0.001, 0.002, 0.003, 0.001, 0.001, 0.001, 0, 0
	9	0.975, 0.007, 0.004, 0.007, 0.001, 0.003, 0.003, 0.001, 0.001, 0, 0, 0
	8	0.975, 0.011, 0.007, 0.001, 0.003, 0.003, 0.001, 0.001, 0, 0, 0, 0
	7	0.982, 0.011, 0.001, 0.003, 0.003, 0.001, 0.001, 0, 0, 0, 0, 0
	6	0.984, 0.011, 0.001, 0.003, 0.001, 0.001, 0, 0, 0, 0, 0, 0
	5	0.984, 0.011, 0.001, 0.004, 0.001, 0, 0, 0, 0, 0, 0, 0
	4	0.984, 0.011, 0.001, 0.004, 0, 0, 0, 0, 0, 0, 0, 0
	3	0.996, 0.001, 0.004, 0, 0, 0, 0, 0, 0, 0, 0, 0
	2	0.996, 0.004, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0

Test Linear Regression Peak and Off-Peak for by range standardised Interdrop and APN/m

In order to determine the linear model for the by range standardised attributes we conducted a linear regression. Similar to the not standardised attributes, with a p value of zero there is for the by range standardised interdrop as well as APN/m a significant linear relationship between off-peak and peak. Given the R^2, we can say that 98% of the variance of the peak day can be explained by the off-peak day (see Table_Apx XI-3, Table_Apx XI-4).

The linear model is as follows:
r_o_interdrop = -0.004 + 1.045 * r_p_interdrop
r_o_APN/m = -0.006 + 0.967 * r_p_APN/m

SUMMARY OUTPUT - Interdrop (by range standardised)

Regression Statistics	
Multiple R	0.992295451
R Square	0.984650263
Adjusted R Square	0.984636774
Standard Error	0.012785982
Observations	1140

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	11.93415077	11.93415077	73000.07672	0
Residual	1138	0.186041771	0.000163481		
Total	1139	12.12019254			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	-0.00455075	0.000449117	-10.13265411	3.67402E-23	-0.005431941	-0.003669559	-0.005431941	-0.003669559
X Variable 1	1.044888143	0.003867302	270.1852637	0	1.037300299	1.052475987	1.037300299	1.052475987

TABLE_APX XI-3: LINEAR REGRESSION LINEAR FOR THE BY RANGE STANDARDISED INTERDROP

SUMMARY OUTPUT - APN/m (by range standardised)

Regression Statistics	
Multiple R	0.994530808
R Square	0.989091528
Adjusted R Square	0.989081942
Standard Error	0.010540861
Observations	1140

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	11.46481304	11.46481304	103184.5813	0
Residual	1138	0.126442896	0.00011111		
Total	1139	11.59125594			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	-0.00640691	0.000562136	-11.39744525	1.42716E-28	-0.007509849	-0.005303972	-0.007509849	-0.005303972
X Variable 1	0.967084325	0.003010627	321.2235691	0	0.961177322	0.972991328	0.961177322	0.972991328

TABLE_APX XI-4: OR THE BY RANGE STANDARDISED APN/M

Practical Clustering Approach: Histogram for Deliveries by Foot/Bike

To further cluster the PC5 areas with less than the turning point of foot/bike deliveries which is around 20-40m interdrop, we use the patterns of the scatterplots (see Figure 0.2). Looking at the scatterplot interdrop and APN/km² we can see that the majority of low urbanised PC5 areas (less than 1000APN/km²) have more than 5m interdrop (see Table_Apx XI-5), which can be confirmed by the histogram outcome, where 97% of those PC5 areas have an interdrop of more than 5m.

Bin(Interdrop in m)	*Frequency*	*Cumulative %*
5	5	2.91%
6	3	4.65%
7	4	6.98%
8	7	11.05%
9	8	15.70%
10	14	23.84%
More	131	100.00%

TABLE_APX XI-6: CUMULATIVE % PER M INTERDROP GIVEN AN APN/KM² LESS THAN 1000

In Table_Apx XI-7 and Table_Apx XI-8, we show the distribution of PC5 areas per minor-route given an APN/km² between 3000 and 5000 and above 5000. All PC5 areas above 5000m have a minor-route less than 5m. The majority (97%) of the PC5 areas with an APN/km² between 3000 and 5000 has a minor-route less than 8m.

Bin (minor-route in m)	*Frequency*	*Cumulative %*
5	194	89.40%
6	8	93.09%
7	5	95.39%
8	3	96.77%
10	4	98.62%
More	3	100.00%

TABLE_APX XI-6: CUMULATIVE % PER M MINOR-ROUTE GIVEN AN APN/KM² BETWEEN 3000 AND 5000

Bin (minor-route in m)	*Frequency*	*Cumulative %*
3	104	95.41%
4	2	97.25%
5	3	100.00%
More	0	100.00%

TABLE_APX XI-7: CUMULATIVE % PER M MINOR-ROUTE GIVEN AN APN/KM² ABOVE 5000

Appendix XII Evaluation of the original Cluster Division

In order to assess if the new proposed techniques result in an improvement, we assess the compactness and separation given the original cluster division. In the original cluster division delivery areas (BGs) are seen as cluster objects. The objects are divided in five clusters based on their average APN/km². Within their benchmarking they also differentiate between car, scooter and remaining deliveries (incl. bike, e-bike and foot) per delivery area.

To assess the overall coherence within the delivery area Utrecht, we calculate the SSE (see Formula 17) with x equal to the PC5 areas and c_i equal to the average by range standardised APN/km² given all PC5 areas with means of transportation i (i=car, scooter, remaining) (see Table_Apx XII-2). The resulting SSE is 101.6.

To evaluate if the original scaling of APN/km² and the three categories of means of transportation might be more suitable for the clustering, we apply this on a PC5 area level. This division results in overall 7 clusters. Five according to Table_Apx XII-1 and two additional once because within cluster E we have next to bike, e-bike and foot also car and scooter deliveries. We calculate the SSE (see Formula 17) with x equal to the PC5 areas and c_i equal to the average by range standardised interdrop and APN/km² of each cluster I (see Table_Apx XII-3). The resulting SSE has a value of 87.6.

This high value can be explained by analysing the cluster proportions. Cluster 7 (APN/KM² higher than 1000) contains 76% of all clusters. As already mentioned in Section 1.3 a delivery area contains areas with different household densities. Taking the average would reduce the extreme values. PC5 areas are with around 0,284km² significant smaller than a delivery area (BG Utrecht is around 520km²), therefore the APN/km² is more precise resulting in a significant higher range. As we can see in Figure 6.6 there are PC5 areas up to 10000 APN/km², which for instance contains a multi-storey building complex. Therefore, the original APN/km² scaling is not suitable for the new clustering.

Given that we have multiple clusters, we can also calculate the Silhouette Coefficient given Formula 21. The resulting Silhouette Coefficient is 0.142, which indicates according to IBM (2016) a poor performance. Therefore, we can assume that the clusters are not well separated and coherent.

TABLE_APX XII-1: CLUSTER DIVISIONS OF THE ORIGINAL CLUSTERING

Cluster	APN/km2	Name
A	> 1000	Big cities
B	500 - 1000	Highly urban
C	300 - 500	Medium urban
D	175 - 300	Lower urban
E	< 175	Rural

TABLE_APX XII-2: CENTRES AND CLUSTER SIZE BASED ON THE 3 CATEGORIES OF TRANSPORTATION

Cluster	Count of PC5 areas within that cluster	Center given average r_p_APNkm²
Car	21	0.001577389
Scooter	23	0.004017204
Remaining	1097	0.168261035

TABLE_APX XII-3: CLUSTER CENTRE BASED ON AVERAGE BY RANGE STANDARDISED APN/KM² AND INTERDROP

Cluster	APN/km²	Means of transportation	r_p_APN/km²	r_p_interdrop
1	< 175	car	0.001577389	0.697559849
2	< 175	scooter	0.004017204	0.299558588
3	< 175	remaining	0.006268478	0.111247221
4	175 - 300	remaining	0.0168321	0.095438834
5	300 - 500	remaining	0.027033855	0.081306265
6	500 - 1000	remaining	0.051600885	0.05877033
7	> 1000	remaining	0.205599042	0.032980598

TABLE_APX XII-4: CLUSTER PROPORTION

Cluster	Number of objects within the cluster	Percentage of PC5 areas
1	21	2%
2	23	2%
3	63	6%
4	30	3%
5	48	4%
6	92	8%
7	864	76%

Appendix XIII PC5 Areas within a Team or Delivery Tour

As the delivery time is not given per PC5 areas, but only on team or postman level, we analyse how many PC5 areas a team or a postman within a tour has and in how many different clusters they would be. The delivery time of deliveries by scooter and car are booked separately and therefore those two clusters (interdrop > 50-70m and >100-120m) are not considered. We know for each team the delivery tours and for each delivery tours the exact PC5 areas which a tour covers. By that we are able to determine the number of clusters per team and per delivery tour. Listing all 1126 delivery tours would make the table unclear, therefore we listed only the average (see Table_Apx XIII-). The number of tour per team are presented in Table_Apx XIII-1

Team	Number of PC5 areas	Number of Clusters	Dominating Cluster (=majority share)
BT-3401Q	29	3	5
BT-3404Q	36	5	5
BT-3421Q	28	5	5
BT-3431Q	49	5	5
BT-3435Q	24	4	5
BT-3437K	31	4	5
BT-3442D	18	4	5
BT-3443U	16	3	5
BT-3445Z	8	4	5
BT-3512L	32	4	6
BT-3514C	26	3	6
BT-3522B	24	4	6
BT-3523J	39	5	6
BT-3526V	29	5	5
BT-3532V	54	5	5
BT-3544N	16	4	5
BT-3552T	13	4	6
BT-3563B	53	4	5
BT-3571L	42	4	5
BT-3582R	52	4	5,6
BT-3605J	55	4	5
BT-3612B	42	4	4
BT-3645G	42	3	5
BT-3702B	45	4	5
BT-3706T	30	5	5
BT-3723B	24	4	5
BT-3734H	46	5	5
BT-3992P	12	2	5
BT-3993D	26	3	5
BT-UT-AU01	6	3	3
BT-UT-AU02	1	1	3
Average	30.58065	3.903226	

TABLE_APX XIII-1: NUMBER OF PC5 AREAS AND CLUSTERS ON TEAM LEVEL AND THEIR DOMINATING CLUSTER (EXCLUDING CLUSTER OF SCOOTER AND CAR DELIVERIES)

Delivery tour	Number of PC5 areas	Number of Clusters
Average	2.23	1.764706

TABLE_APX XIII-2: AVERAGE NUMBER OF PC5 AREAS AND CLUSTERS PER DELIVERY TOUR

Appendix XIV Example for the performance measure record sheet

In Section 3.2 we conclude that the Benchmarking model has to be clear and easy to understand. To ensure that all performance measures are well defined, we advise to fill-out for each performance measure the performance measure record sheet (Neely et al., 1997). In Table_Apx XIV-1 we provide the performance measure record sheet of the performance measure delivery time per mail volume as an example.

Element		Comment & Recommendation
Title	Delivery time per mail item	
Purpose	Compare the internal efficiency by improving the (estimated) mail delivery-time	
Relates to	The process strategy of increasing the cost efficiency by optimising the estimated delivery time	The current Strategy is defined in Section 3.1, but to ensure commitment show the process manager the link to the KPI of the MJ dashboard before implementing it.
Target	lowest score within the cluster	
Formula	$$p_{Cj,Aj} = \frac{\sum_{i \epsilon Cj,Aj} d_i}{\sum_{V=1}^{V=3} \sum_{T=1}^{3} \sum_{i \epsilon C1} (w_{VT} v_{Vi})}$$ $$with$$ $p_{Cj,Aj}$ $= delivery\ time\ per\ item\ for\ cluster\ Cj\ and\ \ Delivery\ a$ $d_i = total\ delivery\ time\ of\ PC5\ area\ i$ w_{VT} $= weighting\ factor\ for\ mail\ sort\ V\ given\ means\ of\ tra$ v_{Vi} = amount of mail of mail sort V within PC5 area i	The weighting factor can be based on the ones given at the costing and economics department (see Table 5.4). In 2018 PostNL can differentiate in their database between mailbox and ring-packages. Until then we suggest to take one measure and weighting factor for both.
Frequency	Once per season	Define the intervals of the measurement based on the volume development within the year. As soon as the volume development changes extremely (for instance Christmas period) start a new interval. By this a better comparability within and between intervals is possible.
Who measures?	Control department	
Source of data	Network Volume Registration (NVR)	

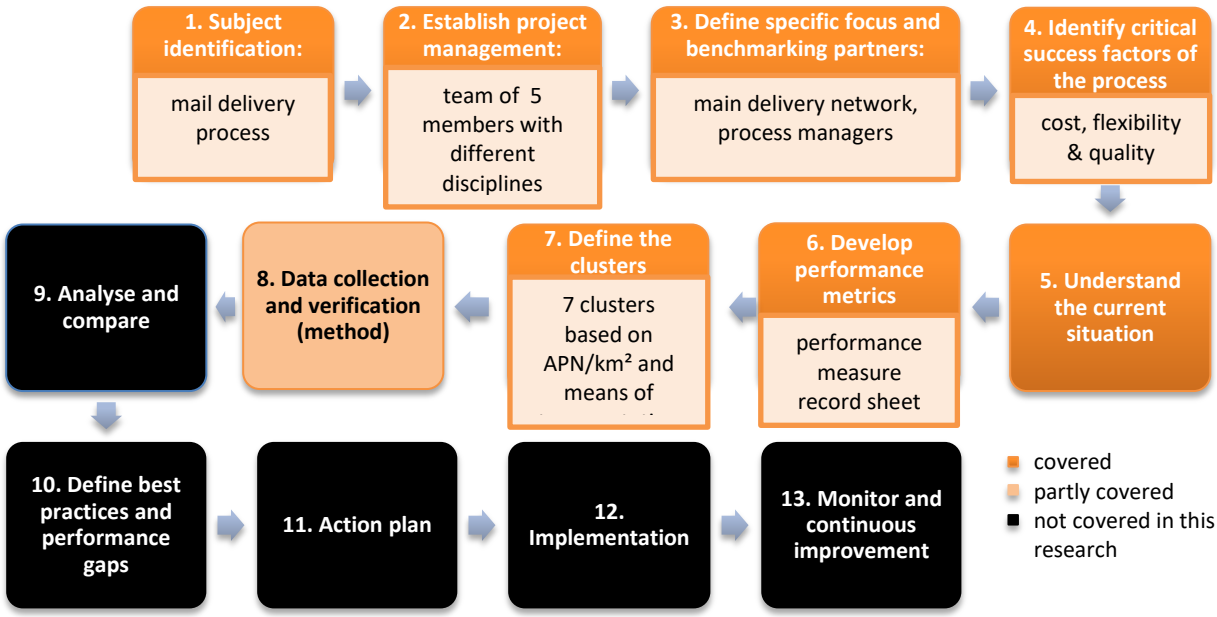
Who acts on the data?	Process manager, team leaders	
What do they do?	Compare the scores within one cluster. In case of abnormalities, zoom in by comparing the data in order to analyse and understand the score difference.	Define together with the process managers action plans and monitoring techniques to ensure continuous improvement. (see Table 4.1: Benchmarking model process)
Notes and comments		

TABLE_APX XIV-1: PERFORMANCE MEASURE RECORD SHEET FOR THE PERFORMANCE MEASURE DELIVERY TIME PER MAIL ITEM

Appendix XV Outline of the Steps required to implement the whole Benchmarking Process

In Section 4.1 we defined the 13 steps of the benchmarking process. At the beginning of Chapter 5, we specified that we only cover Step 1 up and including Step 7 and Step 8 only partly as we only collect the data of delivery area Utrecht. In Figure_Apx XV-1 we briefly summarised the 13 steps and show the findings of the first seven.

To ensure a successful implementation we present in Table 7.1 the tasks, people responsible and time to cover those steps. In the following we outline this table in more detail.



FIGURE_APX XV-1: BENCHMARKING PROCESS FOR POSTNL

Step 7: We defined the clusters only for one performance measure. Thus, for the implementation PostNL has to develop clusters for the remaining performance measures. In Section 5.3 we have shown how to derive from performance measures to cluster attributes, which can be used as a guideline. The senior process manager of optimisation knows all factors influencing the mail delivery process and therefore should cover this task. This task will be time intensive at the beginning (around 80 hours per performance measure), however the results can be used until new performance measures are defined.

Step 8: To calculate the performance measures data of all delivery areas on PC5 level is required. In particular, for the performance measure delivery time per mail volume we first need to collect data on APN/km² and means of transportation (or interdrop) for all PC5 areas within the Netherlands in order to assign them to one of the seven clusters (see Table 6.8 and Table 6.9) (16 hour workload). Subsequently, the data on delivery time and mail volume has to be collected to calculate the performance measure. This task should be covered by the senior controller as she has access to the data set BRPP and NVR. As already mentioned in Section 4.2, performance measures should be rather evaluated based on trends than one snapshot (Criteria 3a). This requires continuous data collection and measurement of the performance. Thus, we advise to automatize the data collection and performance measurement with the help of the IT department.

Step 9 and 10: The analysis and comparison of the performance should be primary done by all process managers individually, but also during the benchmarking sessions under the

guidance of the ambassador of delivery. During the sessions, the derived performance gaps and best practices should be summarised by the ambassador in consultation with the process managers. Secondary, the analysis and comparison should also be done by the senior controller as he has in contrast to the performance managers a more objective perspective and as a controller a better overview of the total performance.

Step 11: An action plan should be defined with the ambassador, the process managers, the senior controller and the process manager of optimisation. The latter can help to define clear actions required to improve the performance or how the best practices can be implemented in the current delivery process. We recommend to design an action plan template specifying the goal, tasks, time frame and the person responsible for the implementation and also the methods for monitoring the best practice and for assessing its success in the long run. This template should be designed by the process manager of optimisation as he knows the essential factors which are required to improve a process.

Step 12: The process managers are responsible of the implementation; however as specified in Chapter 3, even though team leaders and postmen have low power they have to be informed. Team leaders should be informed during their weekly meeting with their process manager.

Step 13: On the short-term, process managers as well as the senior controller should monitor the performance after the implementation to assess if adaptions are required. However, on the long-term after the best practice is fully embedded in the process, the senior controller is sufficient for the supervision. To guarantee the monitoring, we would advise to couple a monitor system directly to the benchmarking model until the best practice is fully embedded in the mail delivery process. If the best practice is critical, PostNL could also consider implementing it into the MJ dashboard.