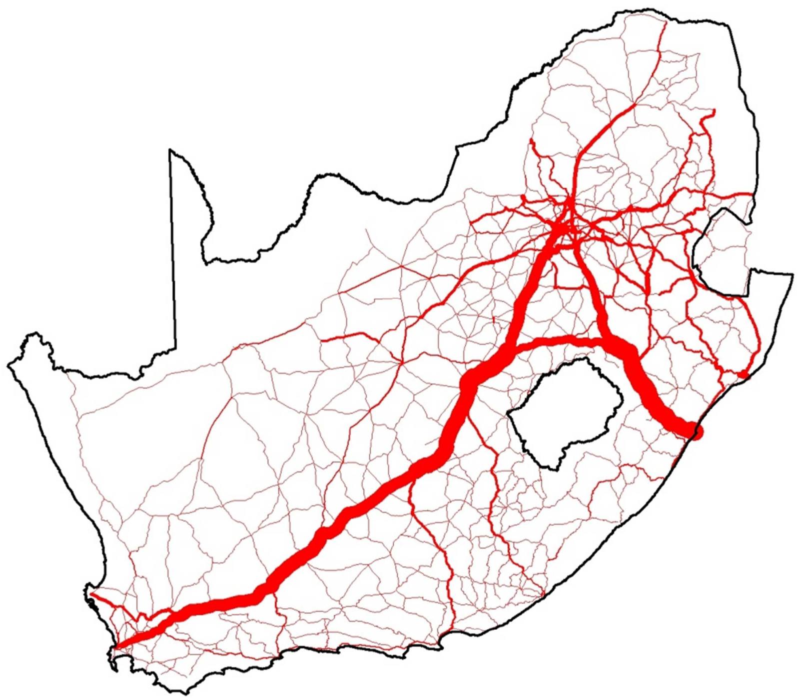
# Decay functions in commodity freight flow models

A regression model for deriving the decay parameter



Author: Adrianus J. Mulder
Date: October 14, 2011
Student ID: S0089621
Academic Institution: University of Twente
Industrial Engineering and Management – Production and Logistics
First internal supervisor: Dr. P.C. Schuur
Second internal supervisor: Dr. ir. L.L.M. van der Wegen
Employer: Centre for Supply Chain Management – Stellenbosch University
External supervisor: Dr. J.H. Havenga

# Management summary

Since 2006 South Africa has her own commodity freight flow model in place that is aiming to predict the freight flows for all goods in the country (aggregated in 64 commodity groupings) up to 30 years into the future. The model is based on gravity modelling; a technique that models the interaction (freight flows) between supply and demand points based on their size and the resistance (for example distance) between them.

An important component in this model, the decay function, describes the decay in volume transported over an increasing resistance variable (distance, time or cost). In general a negative power function or negative exponential function is used. The exponent of the function, the decay parameter, varies for different commodity groupings. It is high for products that are only transported over short distance and low for products that are transported over long distance.

Normally the decay function and its parameter are derived from actual freight flow data obtained from a statistical bureau, a logistical survey or a census. However as no obligation exists for companies in South Africa to hand over freight flow data and the other methods have not yet been carried out in South Africa as they are very costly and time consuming to perform, the current decay functions have been established based on a trial-and-error, lacking scientific background.

Goal of this research has been to develop an alternative way to provide scientific background for and improve the use of the currently applied decay functions for transport by road in South Africa.

The method used is based on regression modelling and assumes that the level of decay can be predicted by an (combination of) other factor(s). Six different factors that were expected to influence decay have been tested in correlation analyses: 'Value per ton', 'Scarcity', 'Supply concentration', 'Local demand', 'Homogeneity' and 'Brand'. The decay in these analyses has been represented by the parameters of negative power decay functions. These functions were derived from actual freight flow data related to 14 different commodity groupings, which were selected based on the amount of gathered freight flow data from industry. This resulted therefore in 14 observations (decay parameter values) for the correlation analyses.

From the expected relationships between the different factors and decay only 'Scarcity' was moderately correlated with decay. The relationships with the rest of the factors were mostly weak and insignificant.

'Scarcity' was therefore the only explanatory variable included in the regression model resulting in the following regression model expression:

Decay parameter = 1.528 -2.628 Scarcity

The explanatory value of this regression model ($R^2$) is 0.479 at a significance level of 0.05. Even at a significance level of 0.003 the relation between 'Scarcity' and decay is expected to be negative.

Because of doubtful data and remarkable differences with the outcome of other studies it is suggested to exclude the commodity grouping 'Bricks' from the regression analysis. This would lead to an increase of $R^2$ to 0.681 with an observed significance level of less than 0.001.

As no verification data are available it is hard to quantify the quality of the model. Given the 95% prediction intervals the accuracy of the model is not very high. Therefore the model can only be used as an indication for the decay parameter value.

Many factors have or may have influenced the correlation and regression analyses. Especially the level of detail (size of the regions), selection of commodity groupings, amount and quality of the obtained data and the processing of the data have had their impact on the values of the decay parameters and scores of the possible influencing factors. Besides the issues with data and methodology also the different time frames (1967 vs. 2010) and countries (US and Europe vs. South Africa) will have influenced the comparison between the studies, the value of the derived decay parameters and their relationships with the possible influencing factors.

The correlation analyses and comparison with similar studies should therefore be made and interpreted with caution.

# Preface

My choice for the assignment was in the beginning solely based on my wish to visit South Africa and experience the country, I fell in love with while travelling, in a different context. However the more I learned about the subject and the influence it has on the logistical developments in South Africa the more interested I became and in the end I spent more weekends figuring out the freight flows in South Africa than I enjoyed her wines and braais (barbeques). Almost turning work into a hobby.

Carrying out my thesis abroad has given some extra challenges in guidance from and communication with my home university but it has on the other hand improved my independence. The whole experience has broadened my horizon academically as well as cultural and social. When I left South Africa I left a beautiful country, a great group of friends and my second home. All in all an opportunity I have to thank both my internal as well as my external supervisors for.

I also want to thank the whole team at the Centre for Supply Chain Management that first of all invited me in their office and gave me a very warm welcome. They always made time to discuss my hypotheses and answer my questions, gave useful advises on different topics and even grabbed the phone to support me in my quest for actual freight flows.

Special thanks are directed to my dear mother who took the time to go through the whole report and was able to point out the obstacles for readers inexperienced with the topic. It has definitely added to the readability of the report.

Yvonne definitely deserves my appreciation for all her help and support. She has been a wonderful friend that always knew how to motivate me but also remind me to place issues regarding my assignment in perspective whenever necessary.

Finally I have to apologize to the contact persons of the different companies that I have called so often to remind them about my request for information that the secretaries started to remember my name.

I believe that although the result of the project might not fully support the initial desired outcome, it has given different useful insights in South African freight flows, resulted in several useful data sets with actual freight flows and a long list of relevant contact persons in the industry. The invitation from the Centre to stay and perform a PhD project was the best appreciation I could receive.

I hope this project will lead to more cooperation between both universities in the future, as there is still a lot to learn from each other.

I am aware of the fact that eighty pages seems a lot but I tried to make the topic understandable for the readers with less experience in the field of gravity modelling without giving in on the depth of the content. Together we will make it till the end.

# Table of Contents

# List of figures

# List of tables

# Glossary

Centre for Supply Chain Management (CSCM)

A consultancy company that is part of the Stellenbosch University and specialized in logistics and change management. CSCM is founder and intellectual owner of the national freight flow model of South Africa as well as the related commodity freight flow model and cost model.

Commodities

Commodities are goods that are assumed to be indistinguishable based on their characteristics important in transportation; value per ton, weight per volume and handling characteristics.

Commodity freight flow model

The gravity based model that is responsible for mapping the commodity freight flows in South Africa from the origins to the destinations.

Commodity grouping

A group of commodities that has been combined to limit the amount of commodities in the model or because the total transport volume of single commodities was insignificant. A commodity group can still be one single commodity (like 'maize') or a combination of commodities (like 'other agriculture').

Conningarth Economists (Conningarth)

A multi-disciplinary economic consulting firm specialised in macroeconomic and microeconomic analysis and econometric modelling in various fields. They establish the necessary OD-matrices as input for the freight flow model.

Cost model

Model, related to both the national and commodity freight flow model, used for the calculation of several logistical cost related indicators. For example the cost related to the truck or train emissions.

Decay function

The function representing the decay in tons transported over a certain increasing 'cost' (distance, time, disutility's). Commonly this function takes the form of a power function ($C_{ij}^{-\beta}$) or an exponential function $\exp(-\beta * C_{ij})$.

Decay parameter

The decay parameter ($\beta$) indicates the slope of the decay function. The lower the decay parameter the further goods will on average be transported.

Distance

The distance from an origin to a destination is in the commodity freight flow model and this research derived from the actual South African road network, penalized for the type of road.

Doubly constrained gravity model

A gravity model in which both the outflows of origins and inflows of destinations are known but it is unknown how the freight is distributed over the origin-destination pairs.

Four step model (FSM)

The four-step model (introduced in the early 1950s), initially meant for the modelling of travel behaviour, has been used for (urban) transportation planning, environmental concerns and multimodal planning during the last few decades. The four steps of the model are: trip generation, trip distribution, mode choice and route choice. They are explained in Section 2.1.

Freight flow

The flow of goods from a certain origin to a certain destination.

Gravity model (GM)

A model derived from physics and currently used in a broad range of scientific fields for example for the trip distribution (second step of the four step model) in freight transportation planning. In this field it describes how goods will flow from origins to destinations following a certain trip length frequency distribution.

Influencing factors (potential)

Factors that (potentially) influence the decay functions (and specifically the decay parameters) in the commodity freight flow model. Influencing factors are for example average value per ton of a certain commodity grouping or the spread in supply points per commodity grouping. They will be referred to as explanatory variables in the correlation and regression analysis.

Magisterial district (MD)

A district ruled by a separate local government. South Africa is divided in about 356 magisterial districts, which are indicated as the origins and destinations in the freight flow models used by CSCM. Every MD in the model has a certain level of supply (production) and a certain level of demand (consumption) for every commodity grouping separately.

Mean trip length (MTL) or Average Travel Distance (ATD)

The average distance a certain commodity grouping is being transported when taking the total set of freight flows for that commodity grouping into account.

National freight flow model

A model that maps all the flows from certain origins to certain destinations on an aggregated level. In this model the freight flows of all commodity groupings from certain origins to certain destinations are accumulated as being one flow.

Origin/destination table (OD-table)

A table with all the freight flows (in tons) from certain origins to certain destinations. Origins and destinations represent respectively the geographic starting and ending point of a freight shipment. They do, within this report, refer to a magisterial district (MD).

Stock Keeping Unit (SKU)

A unique identifier for each distinct product and service that can be purchased. So each SKU refers to a unique item and variants of one product will therefore be considered as separate, unique SKU's.

Sub-commodity

A single commodity within a commodity grouping.

Trip length frequency distribution

A graph that displays the actual or expected distribution of trip length (x-axis) frequencies (y-axis). The distribution of actual trip length frequencies can for example be used to measure the goodness-of-fit of a freight flow forecasting model.

# 1. Introduction

## 1.1 Introduction

This report forms the framework for my master thesis, the final research for my study as industrial engineer at Twente University. The research has been conducted at the Centre for Supply Chain Management, Stellenbosch University, South Africa.

Although there were many earlier attempts, only since 2006 a freight flow model on national level has been in place in South Africa. The model is still evolving through time as new insights and functionalities are implemented. Currently it aims to forecast the freight flows within South Africa 30 years into the future. The model is therefore used to answer strategic logistical and infrastructural questions from a broad range of companies and governmental entities.

For a better understanding of the national freight flows and their behaviour in the future a refinement of the model has been made to a commodity level. The commodity freight flow model is able to predict the freight flows of 64 individual commodity groupings. The input data for this model come from various sources and are updated and verified on a yearly basis. This commodity freight flow model is based on gravity modelling, a technique originally derived from physics. The principle behind gravity modelling tells us that more interaction (trade) takes place between two places if the transport resistance between the places is low. Therefore one of the important elements used in gravity modelling is the component describing the level of resistance of transport between point A and B; the decay function. Instead of the level of resistance level the term attraction value is often used in literature, they are inversely interchangeable.

In case of a pure commodity (for example salt) for which the price of the raw material is more or less equal everywhere, no branding exist and no variation in terms of grades can be determined, the only factor influencing the level of resistance is the distance (transport cost) between the origin and the destination. Because buyers intent to strive for the lowest price salt will always flow from the nearest source to the attracting destination.

By definition a commodity is supposed to be pure and behave accordingly from a freight flow point of view as described above. Purity however would imply defining a commodity for every single SKU (Stock Keeping Unit). This level of data detail is not available and would be impractical for predicting flows on a macro level. In the model commodities are therefore grouped together in 64 commodity groupings. Some of these groups consist of only one product others consist of many products and product groups (often called sub-commodities). The grouping is based on comparable logistical characteristics of the commodities and the total volumes of the individual commodities that are being transported within South Africa.

In case the commodity grouping is not really pure (for example textile) other factors besides distance are expected to influence the level of resistance too. For example brand considerations, value of the good, grades or sub-qualities or the spread of the distribution points.

One way to improve the current mapping and prediction of future commodity freight flows could be by a better understanding of the factors influencing the level of resistance. This is the main topic of this study. In the following section first a more thorough background regarding this topic is given.

## 1.2 Background

The commodity freight flow model for South Africa, introduced in the previous section, was designed in 2006 by the Centre for Supply Chain Management (CSCM) of the Stellenbosch University as a refinement of the national freight flow model. The main purpose of the model is to predict the freight flows between all supply and demand points for the different commodity groupings within South Africa. This process is based on gravity modelling. As the name already suggests is the model based on the law of gravitation; the amount of interaction between two masses depends on their size and the distance between them. In case of freight transportation this means that an increase in the level of supply in a supply point as well as an increase in the level of demand in a demand point or a decrease in the resistance level between the two points will result in an increase in the amount of trade between the two points.

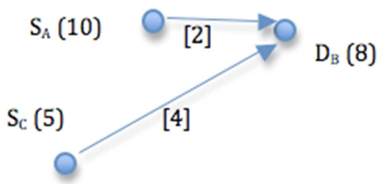
The level of resistance, briefly mentioned in the previous section, can be measured by different metrics: the travel distance from the supply point to the demand point, the travel time from the supply point to the demand point or a general cost factor related to the transportation from point of supply to point of demand. The choice of the metric depends on the purpose of the model and the available data. This will be explained in more detail in the next chapter. Although the level of resistance and the term attraction value are inversely interchangeable we will just use resistance within this report.

As may have become clear from Section 1.1 gravity modelling plays an important role in connecting supply and demand points for impure commodity groupings, as the demand for those goods does not necessarily gets supplied by the closest supply point.

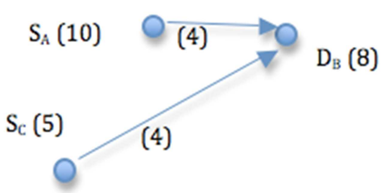
In the next example we will show the consequence of the impurity of commodity groupings and the role of gravity modelling.

Example 1. Dealing with impurity in gravity modelling

In supply point A 10 tons of textile are produced, in supply point C 5 tons of textile are produced and in demand point B 8 tons of textile are demanded. The level of resistance is given by the distance (shown by [x] on the arrows) from supply to demand point.




As the distance from A to B is shorter than from C to B and point A has sufficient resources to fulfil all demand from point B a flow of 8 tons textile is expected to go from A to B according to optimal cost logic. However point A is just producing jeans, point C is just producing sweaters and point B demands 4 tons of both. Therefore in practice A will supply B with 4 tons of textile and C will also supply B with 4 tons of textile (shown by (x) on the arrow).



Because of our aggregation of both jeans and sweaters into the commodity grouping 'textile' the flows of textile do not make sense from a logical (optimal cost) point of view. However in a gravity model these kinds of flows are allowed and modelled, as we will show in the next section. □

As mentioned in the Introduction and shown in Example 1 other factors besides distance like impurity (later represented by the level of homogeneity) are expected to influence the level of resistance too. One way to improve the current mapping and prediction of future commodity freight flows could be by developing a better understanding of the factors influencing the level of resistance. But before we concentrate on the factors influencing the resistance we will first discuss the resistance itself in more detail.

*Decay functions*

The level of resistance is represented by a certain metric (distance, time or general costs) in combination with a decay function. This function represents the decay in volume (tons) transported over increasing distance, time or cost. In gravity modelling the decay function often has the form of a negative power function ($d_{ij}^{-\beta}$) or a negative exponential function (exp ($-\beta * d_{ij}$)). In these functions $d_{ij}$ represents the resistance variable (for example distance) from supply to demand point. Parameter $\beta$ is known as the decay parameter and determines the slope of the decay function. The graph below demonstrates an example for both decay functions.

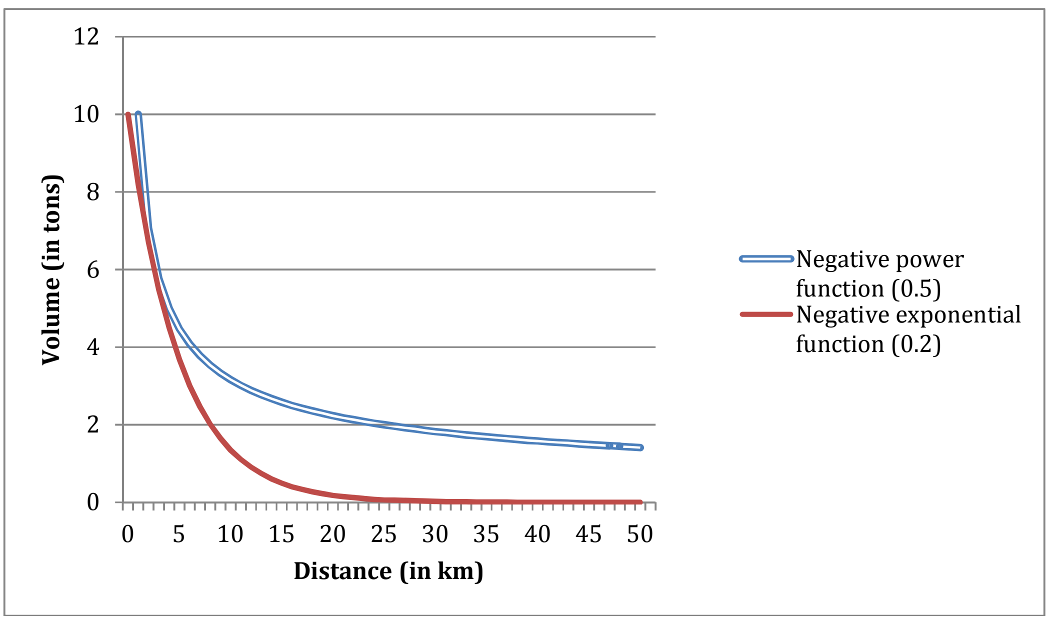


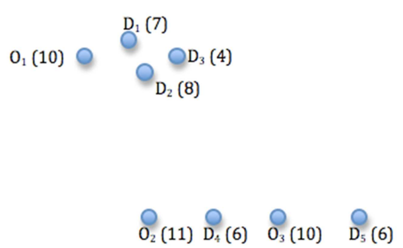

Figure 1.1: Example of negative power function and negative exponential function

The decay function (including the value of the decay parameter is normally estimated from sample data (actual freight flows). The source of the data can be a census, a procedure of systematically acquiring and recording information about a certain group of truck drivers, a logistical survey, a one-time survey under companies that have trucks for transport on road, or a statistical bureau, a governmental institution that gathers all types of data. The estimation of the decay function is generally an iterative process based on the volumes transported and related distances between the supply and demand points. An alternative procedure is based on the mean trip length (MTL) of a commodity

grouping (average weighted distance of all shipments) and the distances between the supply and demand points. However both procedures depend on actual freight flow data and these are not available at the moment in South Africa as no census or survey has been carried out recently. The statistical bureau does not has these data available either, as companies in South Africa are not obliged to hand over information about their freight flows.
The current decay functions are therefore based on a (sophisticated) trial-and-error method that partly rely on decay functions from rail transport, actual flow figures from road transport and knowledge of industry experts. These figures are verified by common sense and feedback from practice. Based on these decay functions the flows are modelled and forecasts are made.
To obtain national wide data for all industries in South Africa necessary for deriving accurate decay functions extensive logistical surveys, census or a new law (obligation to hand over freight flow information) are essential. However both surveys and census are expensive and highly time consuming and the implementation of such a new law is unlikely (because of serious confidentiality issues). Therefore obtaining the necessary data is almost impossible without governmental support or a financial sponsor.

*Flowmap*
Since the introduction of the national freight flow model a specific software package, Flowmap, is used to support the modelling of the flows. The software was designed by the University of Utrecht, the Netherlands, in 1990. It uses supply and demand tables (amount of supply/demand in each point), distance tables (from supply to demand points) and the decay functions as input to generate flows from supply to demand points. The flows are generated in such a way that these flows match the given decay function. This procedure is demonstrated in the example below.

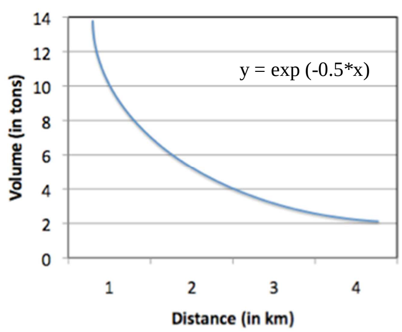
Example 2. Flow modelling in Flowmap
As mentioned before Flowmap uses the supply (in the origins, $O_i$) and demand (in the destinations, $D_j$) (Figure 1.2a), distance from the origin to the destination (Figure 1.2b) and the decay function (including the decay parameter) (Figure 1.2c) as input for the flow modelling.




(a)

| Distance table | |
|---|---|
| $O_1 - D_1$ | 1 |
| $O_1 - D_2$ | 2 |
| $O_1 - D_3$ | 3 |
| $O_1 - D_4$ | 4 |
| $O_1 - D_5$ | 7 |
| $O_2 - D_1$ | 4 |
| ... - ... | ... |

(b)

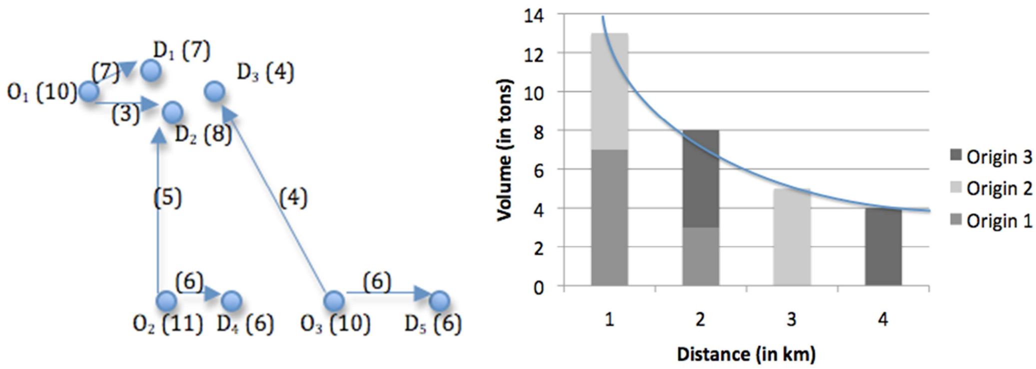



(c)

Figure 1.2: (a) OD – graph; (b) Distance table; and (c) Decay function

In Figure 1.2a the origins ($O_i$) and destinations ($D_j$) and their amount of supply/demand (between brackets) is shown. A complete list of distances from every possible origin to every possible destination is given in the distance table (see simplified example in Figure 1.2b). The graph in Figure 1.2c shows how far the tons of goods are supposed to

be transported. In this case a lot of tons will only be transported over a short distance while just a few tons will be transported over a long distance.
Flowmap will now generate flows from the origins to the destinations. Many different configurations (so-called sets of flows) are possible. Next the different sets of flows (like for example the set presented in Figure 1.3a) are compared with the given decay function (in this case exp (-0.5*$d_{ij}$)).
In Figure 1.3a we can see for example that 7 tons are transported from $O_1$ to $D_1$ and a 3 tons from $O_1$ to $D_2$. The remaining 5 tons of demand in $D_2$ are supplied from $O_2$. The comparison between the set of flows and the decay function is visually represented by Figure 1.3b. Finally the set of flows is chosen that fits best (in terms of distance and tonnage) with the given decay function.




(a) (b)

Figure 1.3: (a) OD – flows; and (b) Flows matched with decay function

The best-fitting procedure within Flowmap is based on a comparison of the MTL from the set of OD-flows and the MTL of the applied decay function (in the range of the shortest and longest OD-distance in the set of OD-flows, see figure 1.3b).
An example of this comparison can be found below (Equation 1.1 and 1.2). In Equation 1.1 $T_{ij}$ represents the total amount (volume) of flows from $O_i$ to $D_j$ (explained in detail in Section 2.2). $C_{ij}$ represents the resistance variable, in this case the distance from $O_i$ to $D_j$, in both equations.
In both equations we multiply the volume with the distance over which it is transported, sum over all these values and divide this sum by the sum of the total volume transported.

*MTL of OD-flows:*

(1.1) $(\Sigma_i \Sigma_j (T_{ij} * C_{ij})) / (\Sigma_i \Sigma_j T_{ij}) = (13*1 + 8*2 + 5*3 + 4*4)/30 = 2$

*MTL of applied decay function* (exp (-0.5*$d_{ij}$))*:*

(1.2) $(\Sigma_i \Sigma_j (exp(-\beta*C_{ij}) * C_{ij} ))) / (\Sigma_i \Sigma_j (exp(-\beta*C_{ij})))$ =
$(exp(-0.5*1) * 1)+(exp(-0.5*2) * 2)+(exp(-0.5*3) * 3)+(exp(-0.5*4) * 4)/$
$(exp(-0.5*1)+exp(-0.5*2)+exp(-0.5*3)+exp(-0.5*4)) = 1.92$

In this case Flowmap compares the outcomes of the two equations (2 and 1.96). If no other generated set of OD-flows produces a MTL closer to 1.92 the flow presented in Figure 1.3a (with a MTL of 2) will be chosen. □

If the decay parameter is unknown Flowmap gives the possibility to use an iterative procedure to derive the parameter based on the MTL of actual observed flows.
In case no applicable decay parameter is available from earlier research and there is a lack of actual observed flows (so no reliable MTL can be derived), as is the case in South Africa, an alternative way to obtain a decay parameter needs to be created.

The overall aim of CSCM is to improve the model every yearly cycle both in functionality as well as accuracy of the current and forecasted flows. As the decay function is one of the main input variables for the flow modelling improving, this function would improve the accuracy of the currently mapped and forecasted flows.

## 1.3 Problem statement

As mentioned earlier the idea within CSCM is that the decay function is influenced by more than just the distance between supply and demand points. The expectation is that a better insight in the factors influencing the decay will improve the accuracy of the decay function and therefore the mapped and forecasted freight flows. This leads to the central research question.

***Central research question***

*How can the current decay functions of the commodity freight flow model applied on the freight flows of South Africa transported by road be improved in a way that both the current mapping of the freight flows as well as the forecasts of the freight flows will be improved by focussing on better insights in the factors influencing decay?*

To answer every aspect of the central research question in a structured way a set of sub research questions are formulated. Combining the outcomes of these sub studies should be sufficient to come to a conclusion regarding the central research question at hand.

***Sub research questions***

*1. How is the commodity freight flow model for South Africa built up?*
*(Which modelling techniques form the backbone of the model and why? What changes have been applied since the introduction of the model? What is the status quo of the model?)*

To be able to speak a common language with the designers and current users of the commodity freight flow model it is important to understand how the model was initially established, which modelling techniques have been applied and what changes the model has gone through. This brings us to the current status of the model. Being aware of the status quo makes it possible to indicate the improvement opportunities of the research on decay functions. Moreover it forms the basis for later validation of the suggested improvements.

*2. Which decay functions are currently used in the commodity freight flow model and why?*
*(What decay function and parameters are applied and why? How are they derived? What is the influence on the model?)*

To make suggestions for improvement understanding the role of the decay functions in the commodity freight flow model is essential.

*3. What decay functions are available/commonly used in scientific research?*
*(What decay function and parameters are applied in comparable models/modelling techniques and why? How are they derived?)*

To be able to question the currently used decay functions it is important to know what other functions are available to describe decay and to know their advantages and limitations.

*4. What factors that could influence the decay in freight transportation are known in literature?*
*(Which are currently applied or have been applied in the past in commodity freight flow models and why? What is the influence on the model?)*

The decay function is currently established based on a trial-and-error method without a proper scientific background. To be able to construct the decay function in an academically appropriate way based on the factors by which it is influenced, these factors obviously need to be known.

*5. How can the present commodity freight flow model be improved by including insights on the factors influencing the decay?*

After our analysis we expect to know how the different factors influence decay and to which extent. These insights should eventually make it possible to establish a regression model that enables CSCM to derive decay functions without performing a nation-wide logistical survey or census or waiting for the implementation of the obligation for companies to hand over their freight flow data. This scientific method should eventually lead to better founded forecasts for the future.

*6. What data are required for the model improvement and how can it be acquired?*

Part of the analyses will be the measurement of the influencing factors and the scoring of the commodity grouping against these factors. This will be the biggest share of data needed and needs to come from external parties. Next to that current flows (output from the model) are needed to test the results of the decay function improvements, data that come from CSCM itself.

*7. How should the improvement of the decay functions be measured and validated?*

Although both the decay functions currently used and the output flows from the model are not proven to be totally correct they are the best values currently available. The validation will be done based on (parts of) the flows that are most reliable.

*8. How should the improved model be used and maintained?*

The improvement will be limited to the decay functions so the changes to the structure of the model are expected to be rather small. The whole model is based on Flowmap. This program will have limitations concerning input variables. That fact will be taken into account throughout the research. The overall perception is that the basic way the model is currently used will not change majorly.

The biggest challenge will be to obtain the necessary input data to score the influencing factors as these scores are likely to change over time. This counts both for the yearly update as well as for the future forecasts.

### 1.4 Research purpose

In the introduction of this report the main purpose of this research, increasing the accuracy of the model output (currently mapped and predicted freight flows) by improving the current decay functions, has already been mentioned. By gathering more knowledge about the factors influencing the decay of transported volume over distance establishing the decay function will become more fact based. Therefore the accuracy of both currently mapped flows and predicted flows is expected to increase. In the current situation the decay function is static for all the predicted years as no indicators for change are available. However Iacono et al. (2008) already indicated that the decay function is often studied over time because of its characteristic of being dynamic and changing in response to transportation network development for example. So by knowing the factors influencing the decay it should be possible to predict changes in the decay function if a change in one or more influencing factors is expected. Therefore a significant increase in accuracy is expected in the predicted flows.

The shift from the current (sophisticated) trial-and-error method towards a more fact based approach to establish decay functions makes it possible to give a validated explanation for the chosen decay function, a requirement of the current clients of CSCM.

An important additional advantage of the fact-based approach is the time consumption of the process. The model has to be rerun for every adjustment of the decay function, which takes a lot of time, especially if it has to be done for 64 commodity groupings individually. The current approach uses trial-and-error, which means a lot of adjustments. The fact-based approach should be able to reduce the amount of necessary adjustments drastically and therefore the time needed for rerunning the model.

### 1.5 Research method

The first steps to understand the composition, role and behaviour of decay functions within gravity modelling and the commodity freight flow model in specific are conducting a literature study and interviewing the people involved in the design of the model, the construction of Flowmap and people working with gravity modelling on a regular basis (the experts).
The knowledge about the basics of decay functions forms the basis for the analysis of factors influencing the decay. Again literature and interviews will form the starting point, resulting in a list with possible influencing factors. Based on statistical methods the real influence of and correlation between the influencing factors should become clear. This should generate sufficient input for the final formula replacing the old decay function.
To be able to carry out statistical analyses sufficient and verified data need to be available. Therefore currently available data need to be verified and in-depth research should be done on the characteristics and current flow behaviour of the different commodity groups. Understanding their characteristics and current flow behaviour are essential to be able to make realistic comparisons among them during workshops with

the CSCM team. Therefore different data sources and knowledgeable people will be consulted.

## 1.6 Boundaries

***The research only concerns the transportation of goods on road***

The commodity freight flow model combines the flows of goods (no passengers are included) from all the different available modes in South Africa. However this research only concerns the transportation of goods on road as the flows of goods on the other modes are known based on detailed and reliable data. Road however counts for about 90% of the total tonnage of freight transported and is therefore of high interest for CSCM and her clients.

***The deep dive analysis will be limited to a maximum of 10 commodity groups***

The deep dive analysis will be limited to a maximum of 10 commodity groups that are expected to cover the upper and lower bounds of the rating spectrum for every influencing factor. This will speed up the process of discussing and verifying figures (rates) during the workshops, limit the amount of time needed for this phase and is expected to deliver sufficient information for the next steps in the research. The other commodity groups will be rated and verified outside of the workshops based on comparisons with the in-depth researched commodity groupings and the derived upper and lower bounds, without carrying out in-depth research on them.

***Verification will be based on the latest modelled flows***

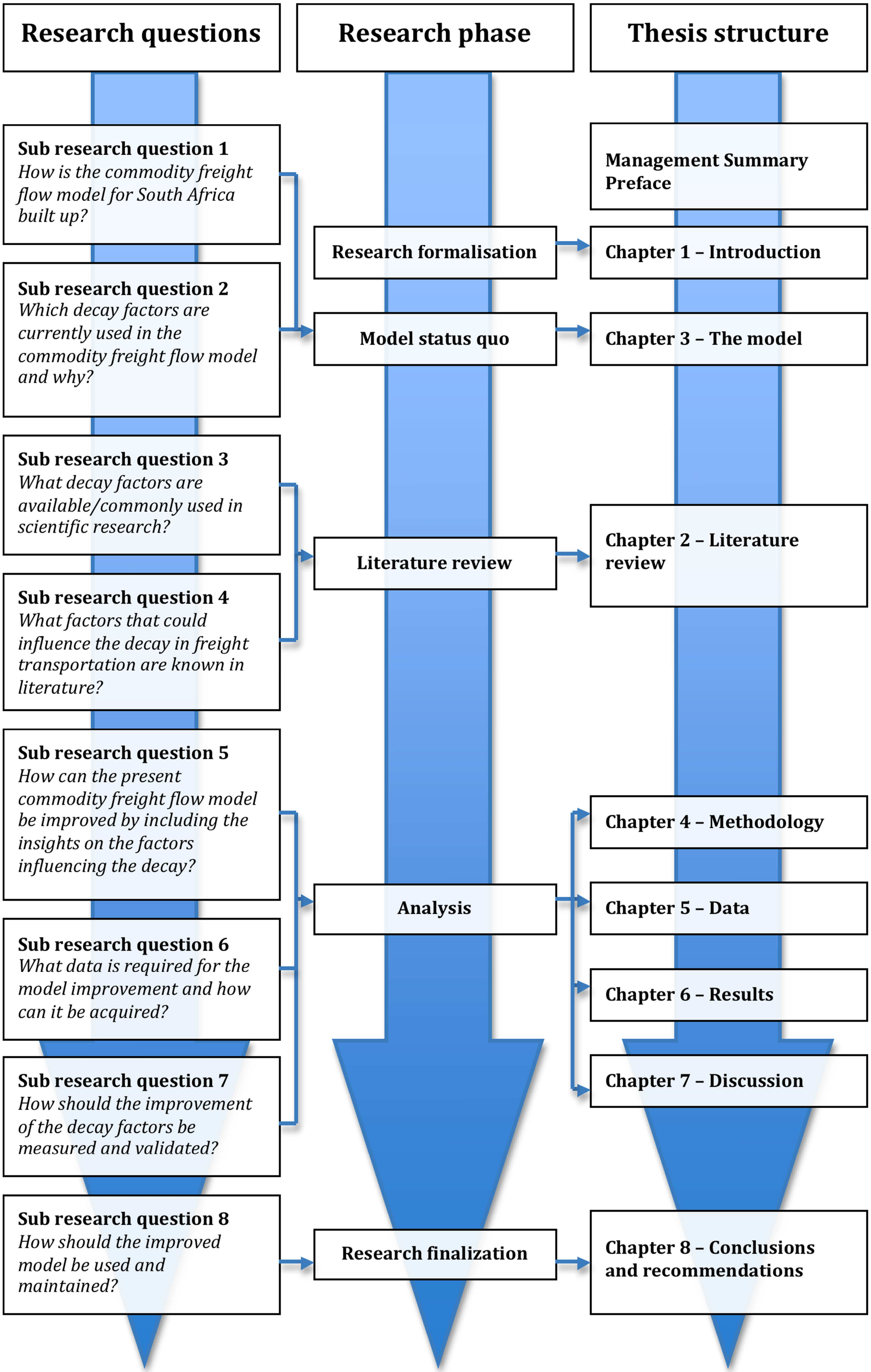
Verification will be based on the latest modelled flows. These flows are no exact replication of the actual flows but they currently give the best indication. The lack of actual flows is a problem experienced often in this field of research as mentioned by De Jong et al. (2004) and Bröcker et al. (2010). In these cases different sources of data have to be combined to establish the overall view. One of the factors influencing the accuracy of the output of freight flow models is the large amount of modifications necessary to make the already scarce, available data suitable as input for the model (De Jong, 2004). Although the modifications may have a serious impact on the output of the model this impact will be the same for both the currently modelled flows as for the flows established based on the model improvements. Therefore the modification issues are not taken into account within this research.

***Manoeuvring within the possibilities and boundaries of Flowmap***

The commodity freight flow model is based on and built up around Flowmap and its required input variables. Changing the software package (switching to another provider than Flowmap) is far from likely as it will be very costly and time consuming. Another possibility would be to change the structure/code of Flowmap. However this will also be a costly and time-consuming exercise, as CSCM does not have the authority nor the necessary knowledge to make changes in the program structure by herself. Therefore the intention of this research is to manoeuvre within the boundaries of what is possible within the structure of Flowmap.

## 1.7 Thesis structure

The remainder of the report is organized as follows. Chapter 2 consists of a literature review where the focus will be on gravity modelling, estimating decay functions, the factors influencing decay and topics from closely related study fields. An in-depth overview of the commodity freight flow model is given in Chapter 3. The analysis methodology is presented in Chapter 4. Chapter 5 describes the data used to perform the proposed analyses. The results of the study are presented in Chapter 6. Chapter 7 is used to discuss the results of the study and Chapter 8 concludes the report with most important findings, an answer to the research question and recommendations for further research. On the next page a schematic overview of the structure is shown.


Research questions
Research phase
Thesis structure
Sub research question 1
How is the commodity freight flow model for South Africa built up?
Sub research question 2
Which decay factors are currently used in the commodity freight flow model and why?
Sub research question 3
What decay factors are available/commonly used in scientific research?
Sub research question 4
What factors that could influence the decay in freight transportation are known in literature?
Sub research question 5
How can the present commodity freight flow model be improved by including the insights on the factors influencing the decay?
Sub research question 6
What data is required for the model improvement and how can it be acquired?
Sub research question 7
How should the improvement of the decay factors be measured and validated?
Sub research question 8
How should the improved model be used and maintained?
Research formalisation
Model status quo
Literature review
Analysis
Research finalization
Management Summary
Preface
Chapter 1 – Introduction
Chapter 3 – The model
Chapter 2 – Literature review
Chapter 4 – Methodology
Chapter 5 – Data
Chapter 6 – Results
Chapter 7 – Discussion
Chapter 8 – Conclusions and recommendations

# 2. Literature review

We will use this chapter to review the history and current application of gravity modelling (Section 2.1). The resistance variables and decay functions as well as the way they are estimated from sample data will be discussed (Section 2.2). We will review the factors known in literature that have been proven to influence the decay functions in freight transportation (Section 2.3) and we present and evaluate decay functions used in comparable models (Section 2.4). We will conclude this chapter by touching on some freight flow related topics (Section 2.5).

In the last decades the decay function has gain a lot of attention in scientific research. Numerous studies especially in travel behaviour of pedestrians (willingness to travel for work for example) and cross border freight transport have been conducted in the last decades, aiming to define and derive the “correct” decay function (Haynes, 1984). However the decay function that should explain the freight flows within country borders seems to be of less interest. The studies that are dedicated to national freight flows only slightly touch the decay function but there seems to be a lack of research questioning distance to be the only factor influencing the decay.
The lack of research on factors influencing the decay in transport freight flows may be partly explained by fact that the field of scientific research on this topic is rather small and the required budget to conduct proper studies in this field is high. Besides that the level of detail and the accuracy of the data available in most other countries with a freight flow model in place are higher than the detail and accuracy of the South African data. Therefore the derived decay functions are more accurate and the need for more information on the influencing factors is lacking (no regression model needed).

The amount of commonly available publications about decay functions within national freight flow models may be limited (Östlund, 2003 and Hensher, 2008) but as said earlier much about decay functions and their role in transportation modelling in general has been researched. We will now first introduce gravity modelling before we focus on decay functions and the factors influencing them.

## 2.1 Gravity modelling

*The universal gravitation law*

Although a theoretical explanation for the application of the empirical gravity equation for commodities was only given in 1979 (by Anderson) the gravity equation was by then already intensively used as a trade device for over a quarter of a century. The principle of the gravity equation, originally derived from the universal gravitation law in physics founded by Newton (in 1687), describes the interaction (force) between two points (masses) based on the size of both points and the distance between them (Roy and Thill, 2004). The level of interaction is proportional to the product of the two masses and inversely proportional to the squared distance between them. The original expression has therefore the following form:

(2.1) $F = G*(m_1*m_2/r^2)$

Here F is the force between the masses, $m_1$ and $m_2$ are the first and second mass and r is the distance between the masses. G is the gravitational constant, a constant of proportionality used to make sure that F has a meaningful magnitude.

The gravity models currently used in the field of transportation are all based on this early gravitation law. As expected these models have for various reasons evolved over time, some just slightly others, like the traffic-demand equations, have started to take on the characteristics of large multiple regression models (explained in Section 4.7) (Taaffe, 1996). Later in this section we will show some of the evolvements in the model expression like the masses that may often be replaced by supply and demand and the squared distance that may be replaced by a combination of a resistance variable and a decay function.

Before the model formulation is discussed in more detail it is necessary to see where the gravity model fits in the broader process of transportation forecasting to understand the model components involved. An easily understood model that is often used for transportation forecasting purposes is the four-step model. The model includes all basic steps involved in the forecasting procedure. We will discuss the model and the separate steps below.

*The four-step model*
The four-step model (FSM) was designed and applied the first time during the early 1950s. It was initially meant for the modelling of travel behaviour to evaluate traffic-engineering improvements. Later it has been used for urban transportation planning, environmental concerns and multimodal (multiple modes of transport) planning (McNally, 2000). Since the late 1970s when was recognised that the initial model was not suitable for the emerging policy concerns many model upgrades and improvements have been conducted. This has led to the start of what has grown to become the activity-based approach and base for the current modelling of passenger transport.
Although the four-step model was, like many other model concepts, created for travel behaviour of passengers initially freight transportation adopted this model too. It has been in place for many years now. Each step had to be adapted to be applicable for freight transportation forecasting but the basic principles have stayed the same (De Jong, 2004):

* Trip generation (supply and demand): the quantities (in tons) of goods to be transported from the various origin zones and the quantities to be transported to the various destinations zones are determined. They form the marginals of the OD-matrices (see Figure 3.3 for an example of an OD-matrix and Sections 3.2.2 and 3.2.3 for more explanation).

* Trip distribution: the flows of goods transported from points of origin to points of destinations are determined, often using a gravity model function. They form the cells of the OD-matrices.

* Mode choice: the allocation of the commodity flows to modes (e.g. road, rail) is determined. This modal model may be of the logit form (statistical model for predicting human choice behaviour), developed by McFadden.

* Route choice: after converting the flows in tons to vehicle-units, they can be assigned to networks.

Although the steps above seem to have a fixed sequence the possibility of feedback is allowed and therefore makes the model principles better aligned with other existing concepts (like supply-demand equilibrium). A drawback of the model is that it has difficulties in capturing adequately the factors that influence shippers and carriers route choice behaviour (Southworth, 2006). Institutional and decision-making structures are simply lacking in current applied models of this kind. These components, appearing in other models, seem to play a significant role in the explanation and forecasting of the flows. However to be able to use these kinds of functionalities certain input data are essential. If these data components are not available the four-step model does not have serious limitations for the user compared to the other models (that obviously need the data to be able to make their extra features usable).

*Recent model expressions*

As mentioned above the gravity model is commonly used in the second step of the four-step model, the trip distribution. However the general model expression for this step differs from the original expression of Newton. The sizes of the supply and demand points are expressed in the level or volume of supply ($O_i$) and level or volume of demand ($D_j$). The squared distance between the points has been replaced by a more general resistance variable ($C_{ij}$) combined with a certain decay function. $C_{ij}$ can still represent the distance between the origins and destinations but also the travel time between these points or general (transportation) costs. The decay function (in this case a negative power function) describes the decay in tons transported versus an increasing resistance variable. Parameter ($\beta$), the decay parameter, determines the slope of the decay function.

The expression may therefore have the following form:

(2.2) $T_{ij} = \alpha * O_i * D_j * C_{ij}^{-\beta}$

In this expression $T_{ij}$ is the amount of trips from $O_i$ to $D_j$ and therefore comparable with the force of interaction (F) in the original model expression of Newton. $\alpha$ is the substitute for the gravitation constant (G) having the same function in reducing the estimated flows to more realistic magnitudes (with $\alpha \geq 0$). The decay parameter ($\beta$) is generally positive. Only in case more volume gets distributed over long distance than over short distance (which is unlikely) the parameter becomes negative.

The gravity model can also be used for a combination of the first and second step of the four-step model, the trip generation and the trip distribution. This model variant is used in case the actual supply and demand values of the different places of origin and destination are unknown (in African countries for example). The variables in the gravity model expression have to be changed again slightly.

If we recall the original gravity model expression, size is expected to be the most important indicator for flow generation. The most commonly used indicator of size is the gross domestic product (GDP), in the new model expression represented by ($Y_i$) for the origin and ($Y_j$) for the destination. Next to GDP other origin and destination specific characteristics (like for example the size of the region of origin, size of population or productivity) can be included (Bröcker, 2010). The expression may have the following form:

(2.3) $T_{ij} = \alpha Y_i * Y_j * C_{ij}^{-\beta}$

This model expression (in more sophisticated form) is often used in cross border freight flow modelling, as GDP is known for most countries with a reasonable level of accuracy.

Many different variants of the original gravity model are currently in use within freight flow modelling and many other research fields. However the basic principles have not changed tremendously. We still see substitutes for the size of origin and destination in every model as well as combinations of resistance variables and decay functions. In the next section we will have a closer look at these last two elements.

### 2.2 Resistance variables and decay functions

In the previous section both the resistance variable and the decay function have been mentioned. The combination of both is referred to as the level of resistance to transport freight from certain origins to certain destinations. The combination determines how far, long or against which costs certain goods are being transported. The role of the resistance variable and the decay function in gravity modelling is significant. A more detailed review is therefore desirable and given below.

*Resistance variables*

In the original model expression of Newton (Equation 2.1, see previous section) the amount of interaction between two masses was inversely proportional to the squared distance between the two masses. Distance can be seen as the resistance variable within this expression. Nowadays distance is still widely used as a resistance variable in gravity modelling however alternatives have been introduced. Instead of transportation distance, resistance can also be measured in transportation time or transportation costs (Iacono et al., 2008). Bröcker et al. (1990) even included geographical, historical and cultural transportation barriers in their resistance variable.

The choice for the resistance variable used in the gravity model highly depends on the purpose of the model and the available data.

In freight flow modelling distance is the most common resistance variable. The necessary data are often more reliable compared to transportation times as only origin, destination and (road) network are needed for the distance (objective data) compared to the travel times which are self-reported and therefore subject to the bounds of human perception and cognition (Iacono, 2008). Moreover from a cost calculation perspective (in case of non-perishable goods) distance is often the most important denominator. The use of travel time as resistance variable is far more popular in urban passenger transportation models, as travel time is the most important denominator for the far majority of the people travelling for example from home to work, school or the shopping mall.

Distance related cost components like petrol and truck depreciation are by far the biggest cost components related to road freight transportation. From a macroeconomic perspective they can be assumed to behave more or less linear with distance. Because of this assumption of linearity the cost components can be represented by distance only and no general transportation cost variable is needed. Model simplicity motivates the preference of distance above a general transportation cost variable in case of road freight transportation. There are situations were certain significant costs cannot be linked to distance or travel time (for example an airplane charter) in these cases a transportation cost variable can be used.

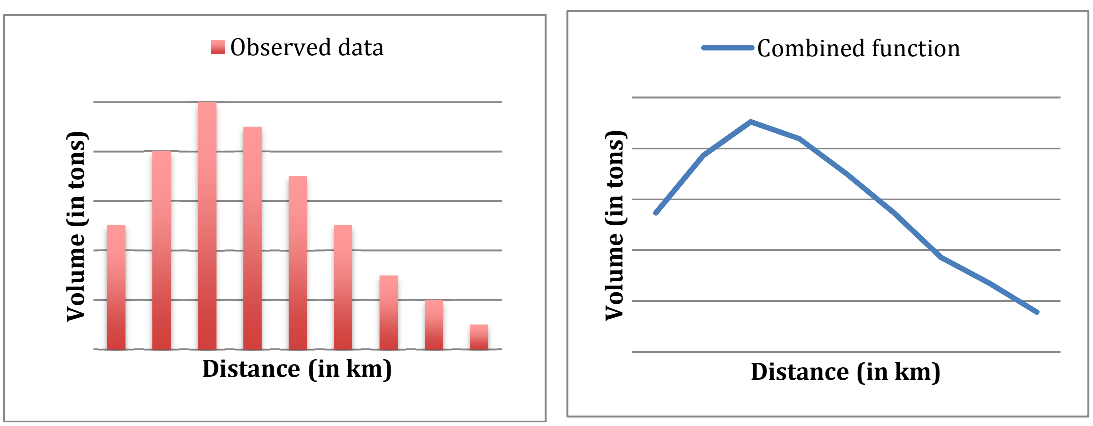
The transportation barriers mentioned by Bröcker et al. (1990) are specifically meant for gravity modelling in international trade.

Next to the fact that different resistance variables can be used there are alternative ways to measure these individual variables.

Distance can for example be based on a real road network, a penalized road network or the Euclidean distance among others. The choice how to measure distance depends on a range of variables but one of the important aspects is the mode of transport as the distance between origin and destination can differ a lot when travelling by airplane following Euclidean distance or by truck using the road network. Another aspect that influences the choice of measurement is the sub-goal of the model, does it aim for shortest distance, avoiding of rural roads, avoiding of congestion etc.
While making these choices and model assumptions regarding the resistance variable it is essential to take the demand for data and the influence on the model complexity into account.

The other component necessary to obtain the level of resistance in gravity modelling is the decay function. This component is reviewed below.

*Decay function*
In the model expression in the previous section (Equation 2.2) the decay function is a negative power function. However the function can also be specified as the inverse of the resistance variable or with the more common specification of a negative exponential function $\exp(-\beta * C_{ij})$. A combination of the functions is also possible. This so-called combined function, $C_{ij}^{\gamma} * \exp(-\beta * C_{ij})$, is often used when the observed data show a pattern comparable to the form of Figure 2.1 (Ortuzar & Willumsen, 2001). This kind of data is for example observed in vehicle transport in urban areas.




(a) (b)
Figure 2.1: Combined function model fit

The type of decay function used in the gravity model depends on the data related to the subject of interaction (sample data). This can be a set of trip lengths of a transported commodity group for example. The decay function that has the best fit with the data, based on a goodness-of-fit test (see Section 4.6), will normally be used. However the choice for the decay function can also be limited by the software used for the modelling

process. This is the case with Flowmap that only provides the choice between a negative power and a negative exponential function.
Regardless of the mathematical expression used, the decay function is intended to convey the decline in interaction as the distance, time or cost between origin and destination increases. As mentioned earlier, the (steepness of the) slope of the decay function is determined by the decay parameter. We give a more detailed description of this parameter below.

*Decay parameter*
Simultaneously with the decay function the decay parameter is estimated from the sample data. In essence there are two primary methods (similar to other spatial interaction models) for estimating a decay function and parameter: linear regression using ordinary least squares on a transformed decay function and estimation of a nonlinear model using maximum likelihood estimation (MLE) techniques.
To be able to use the ordinary least square method the model (for every type of decay function) needs to be transformed into a linear-in-parameters form by taking the natural logarithm of both sides (Fotheringham & O' Kelly, 1989):

(2.4) $\ln(T_{ij}) = \ln(\alpha) + \ln(O_i) + \ln(D_j) - \beta^*\ln(C_{ij})$

The parameters in the model will be unbiased and consistent except for the estimate of $\alpha$, produced as $e^{\ln(\alpha)}$. This constant will be underestimated unless the model fit is perfect (Heien, 1968). Depending on the level of desired accuracy this deviation becomes more important.
The parameters in the unconstrained gravity model can also be estimated via a maximum likelihood technique if it is assumed that interactions are the outcome of a Poisson process (Flowerdew & Aitkin, 1982). Several algorithms for the estimation of the maximum likelihood are available for example the Newton-Raphson procedure (Jennerich, 1976).

These are the two commonly used techniques for estimating decay parameters in case sample data are available from surveys, census or a statistical bureau. The alternative method based on the MTL (see Section 1.2) is less accurate and therefore less frequently used.
In case the sample data are not available, as is the current case in South Africa, we have to look for an alternative, as mentioned in the first chapter. Obviously, as also mentioned in that chapter, CSCM has an alternative way to derive the current decay functions but this method was designed from a practical perspective and lacks academic background. The proposed alternative to this method assumes other factors besides distance to influence decay. Therefore in the next section we will present the findings about these factors from literature.

## 2.3 Factors influencing decay
Repeated applications to real-world transportation situations have made it clear that there is no single “correct” decay exponent (decay parameter) for all situations that reflect some underlying law of human spatial interaction (Taaffe, 1996). Variations in the decay function itself have become the subject of empirical study since it became apparent that decay parameter (β) values will be different for different years, different modes, different commodities transported, to name a few.

Several studies have pointed out evidence for the relation between the value of goods and the decay parameter. It is stated that higher-value goods should be less sensitive to the increased costs of long-haul commodity shipment since transport cost form a smaller part of manufacturer's total costs (Taaffe, 1996). Therefore higher-value goods are associated with lower value decay parameters.
Another negative correlation that is empirically studied is the one between regional specialization in the production of a good and the value of the decay parameter (Black, 1972). In case a region (for example the south-east of South-Africa) is highly specialised in the production of a particular good (for example avocados), it is less likely to be sensitive to distance in its shipments as the avocados have to be transported to the west coast market anyway. A commodity grouping such as stone that is widely available will be dominated by short-haul shipments because of the proximity of competitors and will therefore end up having a high value for its decay parameter. The regional specialization should be seen relative to the other regions examined. So regional specialization is high when there is only one or a few regions able to produce a certain good. Alternatively, if the entire production of the good is consumed locally due to economies of scale, perishability, and so forth, the decay parameter will be high (Black, 1972).
Black also investigated the relation between decay and the value (per unit weight) of goods, however he concluded that regional specialization had a higher reducing effect on the value of the decay parameter than the high-value per unit weight.

Based on the relationships Black had found he tested a multiple regression model to estimate the decay parameter. He included two variables; "regional specialization" (supply side) measured by the proportion of total flows shipped from the largest shipping region and "local consumption" proportion of total flows that are produced and consumed in the same region. This multiple regression (with correlation coefficients for regional specialization of -0.453 and for local consumption of 0.854; both significantly different from zero at the 0.01 level) accounted for approximately 87 percent of the variation in the decay parameters examined.
This method is in line with the alternative approach suggested at the end of the previous section. However for Black the usage of this approach was not motivated by a lack of access to actual flow data, which were available, but was meant as a study to investigate the feasibility and accuracy of this alternative. As he had actual flow data available, he was able to compare the freight flow model output based on the decay parameters estimated by the regression analysis with the freight flow model output based on the decay parameters estimated. For this estimation he used an iterative process based on the actual flow data. Generally, the accuracy of both estimated decay parameters turned out to be nearly similar.

A relationship between the value of the decay parameter and a macroeconomic phenomenon is obviously also possible. For example an increase in the price of petrol will make long-haul shipments less attractive and will therefore result in a stronger decay on the distance goods are transported. Increasing petrol prices will influence the decay parameter of all commodity groupings (as it is a macroeconomic phenomenon). However the relative impact will be less for higher-value goods and is therefore in line with reasoning that higher-value goods have lower decay parameters.

Although the studies mentioned above indicated a great diversity in the used values for the decay function, some scientists have suggested averages. For the decay (negative power function) on road a value of 2.0 has been suggested for the decay parameter based on Chicago interregional data (Helvig, 1964). For a combination of road and rail in Britain a value of 2.5 for the decay parameter was proposed by Taaffe (1996).

From a forecasting perspective the results of several empirical studies concerning the Chicago traffic are also interesting to mention. These studies show evidence for a significant negative relation between time (in years) and the value of the decay parameter (Helvig, 1964). Reasons for this negative relation might be the increasing efficiency of truck shipments and therefore reduced relative transportation costs.

Beside the relevant influencing factors just discussed there are many other characteristics that do show correlation with decay. However most of them turned out to be only related to the origins or destinations (in most cases cities) and not to the commodity groupings. The difference in focus and the fact that the origin and destination effects are already taken into account by Conningarth (see Section 3.1 and 3.2), make these factors currently irrelevant for this research.

The factors that have shown to have significant influence on the decay (value, supply concentration and local demand) in past studies, will be introduced in our research as possible influencing factors in the correlation and regression analysis. Next to these factors other new developed factors will be researched. The inclusion of possible influencing factors and the methodology used to come to a regression model will be explained in Chapter 4.

As we have seen in this section there has been some closely related research carried out in the past. Most of the studies were based on comparable freight flow models in the US and Europe. In the next section we will briefly discuss two related studies in which the used decay functions are presented (which is rather scarce). This should give us an idea of the decay functions and parameters used in other models. The first study is carried out by Black (1971), which formed the base for his experiments with the regression model amongst others. The second study is a recent study performed in the Netherlands regarding the structure of a new to build freight flow model and the initial outcomes (De Jong et al., 2010).
We focus our discussion on the structure of the models, the different decay functions they considered and the derivation of decay functions.

## 2.4 Decay functions in other freight flow models

*Black (1972) in the US*

The gravity model formulation used by Black is comparable to Equation 2.2. The magnitude constant ($\alpha$) has the form of $1/\Sigma(D_j{*}C_{ij}^{-\beta})$, in case of a negative power decay function. In his study it is assumed that the supply and demand in each region is known as well as the flows between the regions. Therefore the decay function (by Black referred to as the friction factor) is the only unknown component in the equation. He uses an iterative procedure to derive these factors for each commodity grouping separately. The procedure begins with a decay parameter with a value zero that is being increased by 0.025 in each step. In each step the model is run and the generated flows are compared with the actual flows. The procedure stops when the correlation between the flows fails to increase. Black only uses power decay functions within his procedure.
The outcome of his study, the derived decay parameter values for both classifications, are listed in Appendix B. Appendix A contains a list of the currently applied decay

functions and parameters in South Africa.

Black uses the 1967 US Census of Transportation for his study. One classification is based on the 24 major shipper groups and the other set is based on 80 commodity groupings (see Appendix B for more detail on the composition of the data sets). Both classifications describe interregional commodity flows between nine regions of the conterminous US. Black excluded flows within the local production area (not defined) and treated the other intraregional flows as interregional flows. The distances between the nine regions are calculated based on Euclidean distance between the centre points of each region. As limitation of his data he mentions that they are based on a dominant element classification. For example the commodity grouping Cotton also contains raw cotton, cotton seed, cotton yarn, etc. this makes it difficult to predict the behaviour of Cotton and compare it with other commodity groupings. This is comparable to the situation in South Africa.

*De Jong et al. (2010) in the Netherlands*
The model is considered as a simple freight flow model that should be able to foresee the government in information on freight flows in the Netherlands, information that is currently lacking. The model will be used for trip distribution and modal split and makes use of several existing econometric and allocation models for the first and last step of the four-step model, the trip generation and route choice (or allocation).
The distribution model is based on gravity modelling and has the main purpose to produce OD-matrices for future years. As input the model uses a base year OD-matrix, total supply ($p_i$) and total demand figures ($q_i$) for every specified zone and a matrix with resistance levels per OD-pair ($r_{ij}$). The used model expression is therefore more or less equal to the model expression in Equation 2.2 (see Section 2.1) except for the fact that there is no proportionality constant included ($\alpha$ has a different meaning here), the resistance variable is OD-pair specific and is therefore not solely related to a commodity grouping and for international relations (origin or destination outside the Netherlands) a border resistance variable is included. This extra variable ($G_{ij}$) is necessary to account for the negative influence of economical, administrative and cultural differences among countries in their international trade flows. If the origin and destination are both located in the Netherlands this resistance variable is zero. The initial decay function is of a combined form and includes both a negative exponential function as well as a negative power function. Therefore the hybrid model expression (adapted to the terminology and representation of variables introduced earlier in this report) looks as follows:

(2.5) $T_{ij} = O_i * D_j * r_{ij}$

with $r_{ij} = \exp(-\beta_1 * (C_{ij} + G_{ij})) * (C_{ij} + G_{ij})^{-\beta_2}$

As the values for $O_i$, $D_j$ and $C_{ij}$ were known, only the values for $\beta_1$, $\beta_2$ and $G_{ij}$ needed to be estimated. This left De Jong et al. with several different options for the final model; the choice between including $\beta_1$ or $\beta_2$ or a combination of them and including or excluding the border resistance variable.

For the estimation procedure the Newton method is used. This method includes a random combination of parameters in the model every time and converges to an optimal solution. The optimal solution is a matrix with estimated flows that fits best with the actual flow matrix. The fit is quantified by the R-squared method.

Based on a comparison of the outcomes of the estimation procedure described above the model is chosen that fitted best with the actual flows. This model only includes $\alpha$ and leaves $\beta_2$ and $G_{ij}$ out of the model. The decay function therefore only exists of a negative exponential function. The related derived decay parameters, one for every commodity grouping (ten in total) that is included in the model, are listed in Appendix B.

Both models are in essence comparable with the model applied in South Africa, as they are based on simple forms of gravity modelling. However the differences in the grouping of commodities (level of aggregation) and the amount and size of the regions used, might give difficulties when comparing the outcome of the models with the applied decay functions in South Africa. Moreover the difference in time period (1967 vs. 2010) and geographically differences between the countries should be considered when a comparison is carried out.

Besides the freight flow forecasting gravity modelling is applied in several other related fields of research. In the next section we take a closer look at the use of this kind of models in passenger transport and cross border freight flow modelling.

## 2.5 Related topics

Flow mapping and forecasting trip distribution in transportation planning models are just two topics from the broad range of contemporary applications of the gravity model. Besides the well-known fields of transportation and physics, the application of the model includes archaeology for examining the location of lost cities (Tobler et al., 1971), biology for describing the spatial spread of plant pathogens (Ferrari, 2006) and economics for understanding (cross border) trade flows and the influence of trade barriers (Bröcker, 2010). However these topics have just a vague relation with the freight flow modelling we are currently looking at. Passenger transport and cross border freight flow modelling have a lot more similarities and are therefore worth a more detailed review.

As mentioned earlier a lot of research has been done on passenger transport and the travel behaviour of pedestrians. And although these studies may give starting points for model improvement in freight transport some important differences between the freight and passenger transport markets need to be indicated to point out the difficulties in translating one model into the other. De Jong et al. (2004) specified the diversity of decision-makers in freight (shippers, carriers, intermediaries, drivers, operators), the diversity of the items being transported (from parcel deliveries with many stops to single bulk shipments of hundred thousands of tons) and the limited availability of data (especially disaggregate data, partly due to confidentiality reasons) as the most important differences between the freight and passenger transport markets.

*Passenger transport*

Despite the differences freight and passenger transport have enough similarities to make a short review of the studies in this field worthwhile. In the field of travel behaviour of pedestrians mode choice and willingness to travel are popular research subjects as it supports (geographical) planning activities and the calculation of measures of accessibility (influence of distance on travel in urban areas modified by the spatial structure of activities). In both topics the role of distance decay is undeniable. However while the concept of distance decay is able to provide a rough proxy for the effect of travel cost on travel decisions, many researchers have noted the incomplete nature of distance as a predictor of spatial choice and travel behaviour. Timmermans (1996) for example has sought to explain disaggregate spatial choices, such as consumers' shopping choice behaviour, in terms of aspatial, general attributes. The results of this and many comparable studies suggest that qualitative, aspatial factors may exert a significant amount of influence on travel behaviour. Nonetheless, given the limited policy tools available to change behavioural patterns, urban planners have stressed the importance of bringing potential destinations into greater proximity of residences, returning the subject to distance.

*Cross border freight flows*

Long distance and cross-border flows grow faster than local and regional flows (Bröcker, 2010). This is one of the reasons interest for cross border freight flows has increased rapidly in the last decade. However there is a long tradition, starting with Linneman (1966), of using GMs in empirical analyses of international (country-to-country) trade flows, which are, of course, closely connected to interregional commodity flow analyses (Celik, 2007). Standard explanatory variables, such as GNP and GNP per capita in both origin and destination countries, the distance between countries, and dummy variables related to adjacency, language, and trading block affiliation have been studied thoroughly and are extended in several more recent studies.

Fotheringham (1989) suggested and studied the influence of competing destination. This variable measures the accessibility of destination j to all (or a subset of) the other destinations. Interaction with destination j decreases, as more competing destinations exist especially when they are close to destination j. Frankel and Wei (1998) created ''overall distance'' variables, which measure how far an exporting or importing country is from all other countries. Use of these remoteness variables is linked to the idea that the remoteness of an exporter from the rest of the world has a positive effect on trade volume. Thus, Frankel and Wei appear to be the first to introduce measures of the spatial structure in international trade GMs (Celik, 2007).

Bröcker et al. (2010) proposed a prediction model for future cross border commodity trade flows. It is a gravity-based model that includes origin and destination specific characteristics (for example GDP, size, productivity), a distance decay function representing the trade impeding effect of transport costs as well as other (trade or language) barriers and a commodity specific elasticity factor. The richer a region of origin (/destination), the more it specializes in (/demand concentrates on) goods that have comparatively high elasticities (luxury goods). Barriers are expected (based on historic figures) to decline over time therefore Bröcker et al. let them approach a lower bound (lowest currently observed barrier) in a smooth manner. The commodity specific elasticities are estimated with a regression method. To avoid misspecifications of the barriers to bias the elasticity estimates the distance decay function and other barriers are represented by a constant effect. This comes at the cost of losing lots of degrees of freedom and therefore renders the standard errors of the elasticities rather high. Eventually the evidence for systematic tendencies in structural change revealed by the estimates appeared to be not very strong.

## 2.6 Chapter summary

Gravity modelling seems to have earned a prominent place in the historic and current field of commodity freight flow modelling. Although the essentials of the model have remained many new variables have entered the model and it has found several new fields of application in the last half a century. Gravity modelling is in different forms still broadly applied for forecasting domestic commodity freight flows. This strengthens the goal of this study to focus on improvements of certain elements in the model rather than proposing the use of a completely new model.

Although the techniques to estimate decay parameters from sample data are validated they cannot be applied in the current situation, as no actual flows are known by CSCM. The alternative of deriving a regression model based on factors influencing the decay has been applied in 1972 in the US. It showed a strong relationship between supply concentration, local demand and decay. Other studies confirm the relationship between decay and the value of the commodity groupings. The inclusion of these three factors among others in the correlation and regression analysis is discussed in Chapter 4.

Based on these validated relations we can state that common decay functions have high value decay parameters for commodity groupings of low value, low supply concentration and high local demand and low value decay parameters for commodity groupings of high value, high supply concentration and low local demand. These common decay functions are either negative exponential functions, a negative power functions or a combination of both.

Many passenger transport and cross border freight flow models have been based on gravity modelling. However the issues that these models address nowadays are not applicable to the commodity freight flow model. Therefore there relevance for this research is limited.

# 3. The model

With the knowledge from the literature study it is now possible to critically review the structure of the commodity freight flow model and related decay functions currently used. After briefly discussing the purpose of the model (Section 3.1) we present an overview of the model structure by looking at the input, the process and the output of the model necessary to fulfil its purpose (Section 3.2). As the derivation of decay functions is of our special interest we will discuss this process separately in Section 3.3. Section 3.4 will be used to reflect on the grouping of commodities within the model. Finally we give an overview of comparable models applied around the globe (Section 3.5).

## 3.1 Purpose

The initial purpose of the national freight flow model was to get insight in the freight flows of all transported goods within South Africa, including import and export and using these insights to answer logistical and infrastructural questions on a strategic level. This initial purpose is still the main driver behind the model. However due to the rapidly growing worldwide interest in environmental issues calculations of for example CO2-emmsions start to play a bigger role in the model. Additional models like the cost model and the commodity freight flow model have been designed and all three models have been updated throughout the years to be able to answer a broader range of questions and to improve the quality and accuracy of the forecasts.

The commodity freight flow model is able to fulfil the second step of the four-step model, the trip distribution (explained in Section 2.1). The first step, generating tables with supply (production) and demand levels for every commodity grouping, is done by an external specialist party, the Conningarth Economists (Conningarth). Although the third and fourth step, mode choice and route choice, are not carried out by the model itself, they are still closely related and a short discussion is necessary for a clear holistic view.
In the next section we will explain the model and her context in more detail focussing on the process of trip distribution.

## 3.2 Structure

As lined out in the previous chapter the trip distribution is normally followed by the mode choice. However in the case of CSCM these steps are performed in reversed order. CSCM has been able to obtain all the actual flow data from all modes except for road (and air, but the percentage of freight transport by this mode is very small and therefore excluded from the analysis). This made it possible to subtract these known flows from the figures in the supply and demand tables. If for example a flow of 10,000 tons fertilizer from Durban to Johannesburg would be known as being transported by rail the supply figure for fertilizer in Durban would be reduced by 10,000 tons as well as the demand figure for fertilizer in Johannesburg. The supply and demand tables that are left after all known flows of non-road modes have been subtracted contain figures that need to be transported by road only. Therefore no mode choice needs to be done after the trip distribution. We will now go through the step of trip distribution in more detail.

### 3.2.1 *Input*

The model needs several different kinds of data to be able to fulfil the first step of trip distribution. The different types of input data are presented and discussed below.

*Supply and demand tables*

The tables with supply and demand figures per magisterial district (MD) for all 64 commodity groupings is obtained from Conningarth. Conningarth is a multi-disciplinary economic consulting firm specialised in macroeconomic and microeconomic analysis and econometric modelling in various fields. The supply and demand tables (two separate tables) they provide may look as follows:

| Magisterial district | DECIDUOUS FRUIT | CITRUS | SUBTROPICAL | VITICULTURE | GRAIN SORGHUM | LIVESTOCK | MAIZE | .... |
|---|---|---|---|---|---|---|---|---|
| **Aberdeen** | - | - | - | - | - | 1.79 | - | |
| **Albert** | - | - | - | - | - | 2.89 | 0.26 | |
| **Alexandria** | - | - | 38.8 | - | - | 7.71 | - | |
| **Aliwal North** | - | - | - | - | - | 2.56 | 0.25 | |
| **Barkly-East** | - | - | - | - | - | 1.65 | - | |
| **Cofimvaba** | - | - | - | - | - | 0.01 | - | |
| **Cala** | - | - | - | - | - | 0.11 | - | |
| **Cradock** | 0.43 | 0.02 | 0.14 | 0.038 | - | 3.98 | - | |
| **Graaff-Reinet** | - | 0.03 | - | - | - | 3.09 | - | |
| **Hankey** | 2.39 | 7.92 | 0.79 | - | - | 1.34 | - | |
| **Hewu** | - | - | - | - | - | 0.11 | - | |
| **Hofmeyer** | 0.69 | - | - | - | - | 1.15 | - | |
| **Humansdorp** | 5.62 | 1.41 | - | - | - | 6.51 | - | |
| **Indwe** | - | - | - | - | - | 0.78 | - | |
| **Jansenville** | - | - | - | - | - | 0.79 | - | |
| **Joubertina** | 8.13 | 4.68 | - | 0.13 | - | 0.18 | - | |
| **Kirkwood** | - | 2.44 | - | - | - | 0.56 | - | |
| **Lady Frere** | - | - | - | - | - | 0.07 | 4.50 | |
| **...** | | | | | | | | |

* All values in ‘000 tons

Figure 3.1: Example of a supply table

Total demand and supply in the established tables are in balance. Subtraction of the total demand (consisting of consumption, investment and export) from the total supply (consisting of production and import) will always equal zero, as the surplus of production will end up being (inventory) investment.

*Actual non-road freight flows*

As mentioned earlier the model needs the actual freight flows from all the non-road modes to be able to subtract them from the initial supply and demand tables. This includes:

* All the freight transport done by rail (including import/export). These data are obtained from the only rail operator in South Africa, Transnet.

* All domestic coastal transport by ships as well as import and export through the seaports. These data are obtained from the shipping line companies operating in South Africa.

* All the movements of liquid freight through pipelines (including import/export). These data are obtained from the pipeline operators.

* All the freight transported by conveyer belts. These data are obtained from the different mining companies active in South Africa.

These data are also used as a check for the figures in the supply and demand tables, which are based on econometric models. If for example the demand level for a certain MD mentioned in the table is less than what has physically been transported into that MD by rail the figures in the demand table need to be adjusted.

*Resistance variable*
One of the input parameters in the gravity model is s mentioned the resistance variable. In the commodity freight flow model distance is currently applied as the resistance variable. Therefore it uses a penalized road network to measure the level of resistance from origins to destinations. The road network is penalized for the quality of the different roads. In the model well-maintained highways are for example favoured compared to dirt roads with a speed limit of 60 km/h.

*Example 3.1 Penalising the road network*
There are two roads, a dirt road (80km) and a high way (90km) connecting city A with city B. Based on only shortest route one would choose the dirt road. But on the dirt road you can only drive 60 km/h, it is bad for the tires, etc. Therefore if the difference is only small, we would prefer using the high way. To make sure that the model chooses the right route we have to penalize the (dirt) roads. If the dirt roads are penalized with a factor 1.2 (and the penalized distance between the cities by dirt road becomes 80*1.2 = 96km) the model will prefer and choose the high way.❑

Based on this penalized road network and the principle of shortest routes a distance table is established that includes one distance for every pair of MD's. This implies that there is no difference between the distance of the route from MD i to MD j and the distance of the route from MD j (back) to MD i. An assumption that is fair from a macroeconomic perspective. A size dependent intra-zonal distance for every MD is included in the distance table (for example the distance from Aberdeen to Aberdeen, see Figure 3.2). Therefore the input distance table may look as follows:

| ORIGIN | DESTINATION | DISTANCE |
|---|---|---|
| Aberdeen | Aberdeen | 35.41 |
| Aberdeen | Albert | 87.32 |
| Aberdeen | Alexandria | 117.85 |
| Aberdeen | Aliwal North | 53.43 |
| Aberdeen | Barkly-East | 118.21 |
| Aberdeen | Cofimvaba | 155.87 |
| Aberdeen | Cala | 89.54 |
| Aberdeen | Cradock | 214.42 |
| Aberdeen | Graaff-Reinet | 289.07 |
| Aberdeen | Hankey | 444.29 |
| ... | ... | ... |

*distance in km
Figure 3.2: Example of a distance table

*Decay functions*
Next to the resistance variable a decay function is required by the model to be able to derive the actual level of resistance. The impact of distance on the amount of tons transported differs significantly among the commodity groupings. This is taken into account by the model by providing every commodity grouping with its own decay function.

As mentioned in the previous chapters the decay function exists of two components, a decay function and a decay parameter. Normally these parameters would be derived from known flows obtained through a survey or census or from a statistical bureau. Unfortunately because of time and financial limitations a less scientific alternative had to be used by CSCM to be able to establish decay functions. In Section 3.3 we will explain more about this alternative procedure. For now it is enough to know that CSCM derives the decay functions and parameters themselves mainly based on rail data obtained from Transnet.

*Truck counts*
SANRAL, monopolist in operating toll roads in South Africa, has truck-counting stations throughout the country. These counting stations record the amount of trucks that pass including their length and weight. CSCM is provided with this information and uses it to check the modelled freight flows. In case the number of trucks counted on specific roads between two cities is far to little to transport all the freight that should pass those roads given the modelled flows, the model (input parameters) or the supply/demand data need to be adjusted.

Now that we have defined all the necessary input data, we can start describing the actual processes that are carried out within the model.

### *3.2.2 Process*

The model itself is responsible for the trip distribution based on the input data. However between obtaining all the data and the actual trip distribution CSCM has to put a lot of effort in making all the data compatible with and applicable in the model. We will now look at the data conversion and trip distribution in more detail.

*Data conversion*
The import/export data from the seaports, that are necessary for Conningarth to be able to establish a complete supply and demand table, are obtained by CSCM in a raw format. All the products transported need to be assigned to one of the defined commodity groupings. If this task is completed Conningarth is able to provide CSCM with a complete set of supply and demand tables for all 64 commodity groupings and 356 MD's.
The next step is to process and convert all the obtained freight flow data from the non-road modes so that these flows can be subtracted from the supply and demand tables. What remains are reduced tables that consist of supply and demand that will exclusively be transported by road.
During this step the figures from Conningarth are as mentioned before checked against the known non-road freight flows.

*Trip distribution – Flowmap*
To execute the actual trip distribution CSCM uses Flowmap, a software program for geographical analyses. We will first present Flowmap in more detail before discussing the actual trip distribution.

Flowmap was designed in 1990 at the University of Utrecht and is specialized in modelling and displaying interaction data. The software is mainly used for passenger transportation but as it is based on gravity modelling it can, without too many adjustments, also be used for freight flow modelling. However at the moment CSCM is the only user that applies Flowmap for this purpose.
At this stage the program does fulfil the needs of CSCM and the limitations of the program have not yet become a serious problem during the trip distribution phase of the

model execution. However the designers of Flowmap are currently looking for a sponsor to turn the program into a dynamic modelling program that would be able to (quickly) simulate the impact of certain infrastructural changes (for example road extensions) on the modelled flows. An interesting feature as different scenarios can be included in the forecasting of freight flows.

The model that CSCM uses within Flowmap to distribute the flows is known as a doubly constrained gravity model. The DCGM is a gravity model in which both the outflows of origins and inflows of destinations are known but it is unknown how the freight is distributed over the origin-destination pairs. In other words the supply (production) levels in the different origins are known as well as the demand levels in the destinations the only thing that is unknown is which origin supplies which destination with how much volume. The sole purpose of the model is therefore to assign a certain volume to each OD-pair in such a way that all demand is fulfilled, supply volumes are not exceeded and the accumulation of all volumes transported over certain distances follows the pattern of a given decay function. This process needs a more detailed description but first we present the gravity model expression for the doubly constrained gravity model as used in Flowmap (Van der Zwan et al., 2005).

1) $T_{ij} = A_i * B_j * O_i * D_j * f(C_{ij}, \beta)$ (trip distribution formula)
2) $A_i = 1 / (\Sigma_j\ B_j * D_j * f(C_{ij}, \beta))$ (origin constraint)
3) $B_j = 1 / (\Sigma_i\ A_i * O_i * f(C_{ij}, \beta))$ (destination constraint)

$f(C_{ij}, \beta) = \exp(-\beta * C_{ij})$ in case of an exponential function
$f(C_{ij}, \beta) = C_{ij}^{-\beta}$ in case of a power function

where:
$T_{ij}$ = the estimated number of trips between origin i and destination j
$A_i$ = the balancing factor for origin i
$B_j$ = the balancing factor for destination j
$O_i$ = the constraint value for origin i, in other words total supply in origin i
$D_j$ = the constraint value for destination j, in other words total demand in destination j
$\beta$ = the decay parameter
$C_{ij}$ = the ‘distance’ between origin i and destination j

The balancing factors ($A_i$ and $B_j$) ensure that the sum of the estimated outflows per origin equals the known origin total supply ($O_i$) and the sum of the estimated inflows per destination equals the known destination total demand ($D_j$). The balancing is based on an iterative procedure, which continues until both factors ($A_i$ and $B_j$) are stable. Depending on the amount of supply points ($O_i$) and demand points ($D_j$) the computation time for the model can differ significantly, from about a few seconds for single supply or single demand cases up to several hours for the commodity groupings with supply and demand points in almost every MD.
Formula 1 calculates the actual trips in the OD-matrix.
Formula 2 takes care of equating the total number of trips from a certain origin in the matrix to the set value.
Formula 3 takes care of equating the total number of trips to a certain destination in the matrix to the set value.

This gravity model expression does not differ much from model expression 2.2 (see Section 2.1) except for the constraints on the origins and destinations. These constraints are not included in model expression 2.2 as that is an unconstraint gravity model (Williams, 1976). In this type of gravity model no limitations to the amount of trips departing from a certain origin or to the amount of trips arriving at a certain destination are set.

*Trip distribution – Actual process*
If a base year with known flows is available Flowmap will start with calculating the mean trip length and the decay functions for all the commodity groupings. If no base year data are available, the user is able to insert own mean trip lengths (optional) and decay functions.
If no base year data are available and no MTL is known from other sources (like an industry survey) Flowmap is able to generate an MTL based on the figures from the supply and demand tables (all possible OD-pairs), the distance table (minimum and maximum possible distance between OD-pairs) and an inserted decay function. This MTL is essential for the next step.
In the next step the model starts the trip distribution by assigning volumes to the different OD-pairs, which implies that the assigned volume of the commodity grouping is transported from a certain origin to a certain destination. These volumes are recorded in an OD-matrix. The rows in the matrix represent the origins and the columns represent the destinations. So all the volumes in one row are coming from one origin and all volumes in one column are transported to one destination. The volumes are assigned in such a way that the weighted average transport distance (combining the volumes in the cells of the OD-matrix with the related cells in the distance table) equals the given MTL. After assigning volumes to each OD-pair the total volumes in each row and each column are compared with the figures in the supply and demand table, the origin and destination constraints. If the difference between the assigned volumes and the constraint volumes is bigger than the pre-set convergence criterion the volumes in the column or row with the biggest deviation are adjusted. This process is repeated until the convergence criterion (manually set) is reached.
This procedure is done for each commodity grouping separately.

The outcome of the trip distribution process is a matrix with assigned volumes transported from each origin to each destination. As mentioned this matrix is called the OD-matrix and will be explained in more detail in the next section.

### 3.2.3 Output

After the process of trip distribution we have obtained an OD-matrix for every commodity grouping separately. An example of such an OD-matrix is shown below.

| **Textile** | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Origin\Destination | **Aberdeen** | **Albert** | **Alexandria** | **Aliwal North** | **Barkly-East** | **Cofimvaba** | **...** | **Total** |
| **Aberdeen** | 0.67 | 0.23 | 0.03 | 0.66 | 0.06 | 0.17 | ... | 36.2 |
| **Albert** | 0.33 | 1.33 | 0.87 | 0.16 | 0.00 | 0.77 | ... | 25.6 |
| **Alexandria** | 0.05 | 0.2 | 0.43 | 0.07 | 0.00 | 0.03 | ... | 7.3 |
| **Aliwal North** | 0.12 | 0.14 | 0.03 | 0.56 | 0.08 | 0.2 | ... | 12.3 |
| **Barkly-East** | 0.38 | 0.00 | 0.00 | 0.53 | 0.00 | 2.67 | ... | 18.0 |
| **Cofimvaba** | 0.00 | 0.00 | 0.00 | 0.21 | 0.00 | 0.36 | ... | 3.5 |
| **...** | ... | ... | ... | ... | ... | ... | ... | ... |
| **Total** | 13.8 | 9.5 | 17.2 | 39.1 | 3.9 | 26.4 | ... | |

*in '000 tons

Figure 3.3: Example of an OD-matrix for the commodity grouping textile

From the example we can see that from the total production of textile in Aberdeen 670 tons are consumed locally, 230 tons are going to Albert, 30 tons are going to Alexandria etc. We can also see that Aberdeen receives 330 tons of textile from Albert and 50 tons from Alexandria.

In the following section we will discuss how this matrix is used in the steps succeeding the trip distribution.

### *3.2.4 Next steps*

As modal choice, described as the third step of the four-step model, is already done prior to the trip distribution route choice is the next logical step in the process of freight transportation forecasting.

*Route choice - Flowmap*

After generating the trip distribution the flows need to be mapped on the road network of South Africa. As mentioned earlier the model does not take any behavioural choices of truck operators in account. Neither does it include capacity restrictions on the shortest routes (resulting in congestion). The generated flows are simply assigned to the shortest route between two MD's, the routes used in the distance tables, even when there are more routes between the OD-pairs possible. Therefore we cannot really speak of a route choice.

The step of route choice is generally executed by a far more advanced and specific traffic model. These models are available but as they concentrate on human behaviour the necessary data to apply this type of model are of a far more detailed level than the gravity model for national freight flows. However from a macroeconomic perspective the shortest route choice seems appropriate.

The mapping is performed by Flowmap but mainly done for rough visualisation purposes as the realism of the flows over the routes is limited due to the reasons mentioned above. An example of a map with plotted freight flows is shown below.

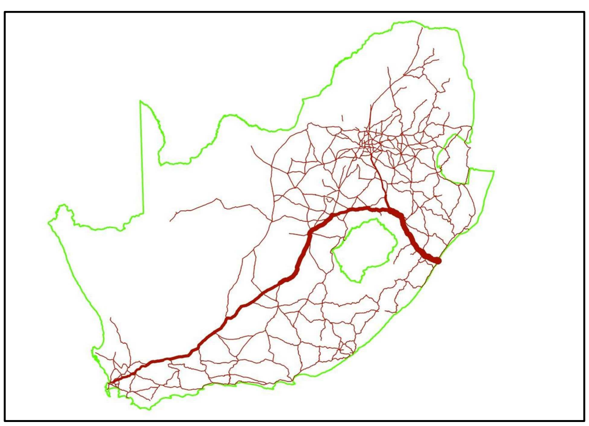


Figure 3.4: An example of a freight flow map for textile

In the freight flow maps (like the example above) the thickness of the displayed road indicates the volume of freight that is transported over it. The lines do not display the directions of the freight flow. Therefore the related data need to be consulted.

After the route choice, the last step in the four-step model, reality checks need to be carried out before the OD-matrix can be used for prediction purposes or answering strategic logistical questions.

*Reality check*

The truck counts, mentioned earlier as being input figures for the model, are not used in the trip distribution generating process itself but for the purpose of performing a reality check on the generated trip distribution afterwards.

As described these truck counts are physical truck counts done along various roads in South Africa. Unfortunately the type of load (commodity grouping) that is being transported by the trucks is not registered but as they do register the weight of all the trucks passing it is possible to calculate (by estimating the weight of the empty truck itself) the total freight volume passing on certain parts of the road network. If the OD-matrices of all commodity groupings are accumulated it gives you the estimation of the total freight flows in South Africa. As both the truck counts and the accumulation of OD-matrices are supposed to give you (an estimation of) the total freight flows on certain parts of the road network it is possible to compare both volumes. In this way the comparison of the volumes works as a reality check on the generated trip distribution.

However as mentioned the value of this reality check is limited because the model only assigns freight flows to the shortest routes between supply and demand points. If for example warehouses are located off-route this may influence the amount of trucks on the shortest route, resulting in strong deviations.

Another reality check is carried out by comparing the generated flows with actual flows obtained from industry studies and expert knowledge. In case a known flow is not generated by the model CSCM will first verify the supply and demand figures from Conningarth. In practise an under- or overestimation of the supply or demand level has always been the cause of the disagreement between the known and generated flows.

It should be clear by now that even with these two high level reality checks in place proper validation of the model is not possible. In general validation of the model is based on a comparison with actual freight flows. However these actual freight flows are not known by CSCM. So currently there is no way to measure the accuracy of the model. We will come back to this point in Sections 4.7 and 6.3.

*Future freight flow predictions*

When the reality checks have been carried out and Conningarth has processed the suggested adjustments (like actual freight flows) for the supply and demand tables, Conningarth is able to generate the prediction based supply and demand tables for the future. These tables are the main input for the predictions of the freight flows generated by CSCM. All input parameters like the decay functions and the road network characteristics are kept the same for all forecasted years as are the percentages of modal split.

*Strategic logistical enquiries*

Both the overview of the current freight flows as well as the future predictions of the flows serve the initial goal of answering strategic logistical enquiries. The type of enquiries as well as the topic can vary a lot within the field of freight logistics and therefore often need more specific information than just the basic flows of certain goods between their origin and destination. To obtain this information a lot of additional tailor-made modelling and (industry) research is often necessary.

This section has made clear that decay functions (and their related MTL's) play an important role in the process of freight transportation forecasting for South Africa and the trip distribution in specific. Although CSCM did not had the luxury of having actual flow data from a survey, census or statistical bureau they have been able to derive decay functions in an alternative way. In the coming section we describe this currently used method for deriving decay functions.

### 3.3 Deriving decay functions

When the model was run for the first time, the trip lengths of freight transported by rail were used as input for Flowmap to establish the distance distribution curve of the different commodity groupings. Decay functions, including the decay parameters, could be derived from these curves. These functions were used as input for the modelling of the flows. In case the flows were not appealing to reality more in-depth research would be done on the flows of the commodity grouping. This could result in adjustments of the decay functions.
For most commodity groupings that were transported both by rail and road, CSCM expected the distances freight was transported by rail to give a reasonably good indication for the distances freight was transported by road. Therefore the decay functions derived from rail data were expected to give also a good indication for the decay functions of transport on road.
Next commodity groupings were compared and groupings with (expected) comparable flow behaviour would be assigned similar decay functions. The similarity was mostly based on industry knowledge (of experts) and common sense.

However problems with this method occurred when large percentages of the commodities were transported by rail, as the behaviour of freight transported by road could differ a lot from the flow behaviour on rail. This is less likely for commodity groupings that have a fairly equal percentage of freight transported on road and rail. Another issue arises from the fact that a lot of commodity groupings are built up out of many sub-commodities. It could be the case that only a certain sub-commodity is transported by rail but the values do not make sense for the rest of the commodity grouping that is transported by road.
Problems also appeared in case the amount of points of source and destination for certain commodities was limited as this limited the amount of different possible trip lengths, making it difficult for Flowmap to come up with a solid (and accurate) decay curve.

In a later stage the decay functions were adjusted based on actual flow data. This was mostly done by a trial and error method of establishing flows that correlated sufficiently with actual flow values found.
Nowadays just minor adjustments are made in the decay functions based on issues that arise from practical flow experience (from users of the model output like Transnet).

Although the method does not seem to be very sophisticated it has some major advantages above the other (historically used) methods; truck counting and transportation surveys. In contrast to the truck counting this model makes it possible to

map the freight flows on a commodity group level. Truck counting, which only counts trucks passing by road, does not give insights in the kind of freight that is transported. This makes it almost impossible to make any substantiated future predictions about the freight flows. However as described in the previous section it can be and currently is used as a check up on the (commodity) freight flow model.
The problem with the survey is the validity of the obtained data, as the response rate is in most cases low and the questions are answered by random people (with a truck), possibly in a subjective way.
The method used for the model in SA serves the purpose of forecasting best because of the detailed level of data on the 64 commodity groupings, making it possible to check the modelled flows in several ways after running the model.

The grouping of goods into commodity groupings plays a crucial part in the way the process of freight transportation forecasting is being modelled. In the next section we will explain how the 64 commodity groupings are established and by whom.

### 3.4 Commodity groupings

Commodity groupings can exist of one (for example maize) or more sub-commodities (for example processed foods). The grouping has been done by CSCM in accordance with the wishes from their biggest client, the only rail operator of South Africa, Transnet. All goods have first been divided in three general classes; agriculture, mining and manufactured goods. Within those classes further break down has taken place based on the total volume (in tons) transported of the different goods. All the goods that have insignificant transportation volume are merged in a group ‘others’ (for example ‘other agriculture). In case certain goods were of special interest because of for example high export or import volumes they are considered as a separate commodity grouping. As mentioned earlier each commodity grouping has been assigned an own decay function. A list of all currently applied commodity groupings and assigned decay functions can be found in Appendix A.
The commodity groupings follow the classification of the widely used SIC codes (Standard Industrial Classification of all Economic Activities) listed for South Africa in the 5$^{th}$ BMR report (Du P Potgieter,1993).
The top 28 commodity groupings (in terms of volume) are more extensively researched by Conningarth than the other 36 remaining commodity groupings. These 28 commodity groupings represent 90 percent of the total freight transported in South Africa.

Now that we have a clear view about the structure of the model and its components we are able to make comparisons with similar models outside South Africa.

### 3.5 Other models

Traffic forecasts are used for several key purposes in transportation policy, planning, and engineering: to calculate the capacity of infrastructure, e.g., how many lanes a bridge should have; to estimate the financial and social viability of projects, e.g., using cost-benefit analysis and social impact assessment; and to calculate environmental impacts, e.g., air pollution and noise. The forecasting models capture and produce an extensive amount and variety of data and therefore allow a broad range of studies to be carried out. This is also a necessity as these studies have to cover the significant investments made to establish the models (Hensher, 2008).

Although the main reasons for constructing the models nowadays may be quite comparable among different countries the available data, followed procedures and detail of model output may differ significantly. In the following section other modern freight flow models will briefly be reviewed but first some earlier South African attempts will be discussed.

*A historical overview of freight flow modelling in SA*

In the last decades several logistical studies in South Africa were (partly) dedicated to mapping the freight flows in the country (Smith, 1971), (Hamilton, 1983), (Pretorius, 1990) and (Freight transport data bank report, 1990). Their main purpose was to get insight in the truck costs and related issues. This difference in study purpose is one of the reasons for the lack of detail in the freight flows itself, which were only part of the input data, and lack of explanation for or reasoning behind the established flows.

The data used as explained earlier came from surveys (with low and subjective response) or truck counting (without information about the specific goods transported) resulting in low reliability and a lack of detail in the output of the models.

Moreover, the way of grouping commodities and defining the origin and destination zones differed between every study. This made it hard to compare the models and their outcomes in a meaningful way.

Although not quantitatively tested the improvements of the current flow model for the purpose of forecasting national freight flows from a qualitative perspective (level of output detail and reliability of data) seem significant. To place the model in a modern perspective a review of other currently operating freight flow models worldwide is required.

*Freight flow modelling around the globe - Europe*

Throughout the world several different types of freight flow models are applied. In Europe alone 65 freight transport models and 29 combined freight and passenger transport models were existing in 2002 (European Model Directory). Underlying principles and necessary input data as well as system functionalities and output detail differ significantly among the different types. The models have been classified in several ways. One of the most recent extensive classifications focussing on the models in Europe has been carried out by De Jong et al. (2004). This classification is based on the earlier mentioned four-step model approach, relating the types of models to the four different steps involved in freight flow forecasting. Within every step De Jong lists relevant currently applied models and refers to recent publications describing their purpose, data requirements, structure and output format.

In the step of trip distribution two types of models are distinguished: gravity models and input-output models. The former type of model is included in the following models among others:

* Dutch TEM-II model (Tavasszy, 1994)
* Dutch SMILE model (Tavasszy et al., 1998)

The Strategic Model for Integrated Logistics and Evaluations (SMILE), one of the first models for decision-making purposes in strategic national freight transport, was designed in 1997 and has since then been used in the Netherlands for several studies and analyses (Tavasszy, 2006). The goal is to perform long-term freight flow forecasting (30 years) and policy impact analyses. Therefore 40 regions and 50 commodity groupings are divined. Similar to the South African model SMILE uses a gravity model for the distribution of trips. Depending on the character and phase of the decisions making

process that needs support the input for the model ranges from economic growth figures to policy scenarios for domestic freight flows. Unlike the South African model SMILE makes a clear distinguish between freight flows generated by domestic production and consumption and the freight flows that cannot directly be related to the Dutch economy. These flows include for example transit flows using the ports for transhipment and intermodal change only (Tavasszy et al., 1998). Both flows are treated separately and in a different way and therefore give unique opportunities for specific policy analyses of both flows.
A more recently designed freight flow model for the Netherlands, GoodTrip, is a logistics-based freight model focusing on the modelling of logistical chains rather than origin/destination data of intermediate products (explained in more detail below). The model also assigns flows to routes based on a shortest route algorithm. However it concentrates on urban freight distribution instead of the macroeconomic focus of the South African model.
As mentioned in Chapter 2 a third model, BasGoed, is currently being developed in the Netherlands. This model is in more extend described in Chapter 2. Both BasGoed and GoodTrip are based on the earlier modules of SMILE.

The input-output models are able to perform both the step of generating supply (production) and demand tables and distributing trips. The national model applied in Italy (Cascetta, 1997) is an example of a multiregional I/O model. The freight flows between OD-pairs follow from a multiregional I/O analysis; region-to-region trade flows for each economic sector using final demand figures and elastic coefficients for inter-regional trade. After converting the value figures into volume figures they are regionalized (divided over the different regions based on for example their population size). The same principle is used in the following models among others:

* STREAMS (Williams, 2002)
* SCENES (SCENES Consortium, 2001)
* NEAC (Chen et al, 2000)
* Fehmarn Belt freight transport model (Fehmarn Belt Traffic Consortium)

A big difference between the two types of models used in the trip distribution is the data requirement. The gravity model is based on relatively simple supply and demand data while the I/O model needs multiregional input-output tables (based on detailed regional input figures). On the other hand the latter models have a strong link to the economy and are able to include land-use interactions (changes in the road network effects the economy and vice versa). The modelling of policy effects is limited for the gravity model. In case of the I/O model it strongly depends on the coefficients used, fixed coefficients reduces the possibilities of modelling of policy effects significantly compared to elastic coefficients.

*Freight flow modelling around the globe - Europe*
A more recent classification of freight flow models was conducted by Yang et al. (2009). This classification focuses on the models applied in the US and uses another approach to classify. The authors distinguish in total 7 classes of increasing complexity and application purposes. The freight flow model of South Africa fit best in class D, the Four-Step Process Commodity Model. Many American states have built and implemented a comparable model. The Wisconsin model is considered a good and complete example as it predicts both the states passenger and truck traffic volumes (Proussaloglou et al., 2007). This class of models is appreciated for its link between the

vehicle flows and commodity flow patterns (Yang et al., 2009). This is not really applicable to the South African model as the link with vehicle flows is currently not present. However the applicability of the models is limited to regional areas as it is not able to realistically capture urban service trips and drayage activities. This latter disadvantage is irrelevant for macroeconomic applications.

*State-of-the-art – logistics based models*

Models of all kinds described above are evolving constantly as new ways of capturing data are developed, new fields of application are indicated and different models are made compatible to one another.

The same counts for SMILE and her derivatives. This model was the first to integrate trade flows and transport flows in one module whereby alternative chain types and warehouse locations are considered using the characteristics of products, industries and transport services (Östlund et al., 2003). This kind of model is called logistics based freight model as it establishes logistical chains by connecting activities from consumers (via supermarkets, hypermarkets and distribution centres) back to producers. They also take into account how consolidation of shipments influences shipment sizes, costs and mode choice. All in all are the estimates of transport flows as outcome of these models more accurate compared to for example the South African model as the costs of transporting goods over a certain geographic distance includes the inventory\ holding costs and handling costs at warehouses (Östlund et al., 2003). Another advantage of the supply chain approach is that each OD-relation in the supply and demand table can be transformed into a chain of multiple OD-pairs, each using different inventory locations. In that way the geographical position of each region in the chain is taken into account when it comes to modelling for example the use of warehouses.

Norway and Sweden have recently adopted and implemented the idea to insert a logistic module into their currently existing freight flow models. It focuses on capturing individual shipments instead of customers or shippers. This approach is discussed in De Jong and Ben-Akiva (2007). This development in model innovation might be interesting for South Africa given their wish to include the different logistic cost factors in their model.

## 3.6 Chapter summary

The commodity freight flow model, aiming to predict the freight flows of 64 commodity groupings 30 years into the future, is used to answer strategic questions of public and private enterprises in South Africa. One of the steps of the model, the distribution of the trips is based on gravity modelling and the supply and demand data of an external party, Conningarth. The available data from other parties make it possible to distinguish road from all other transportation modes. Therefore there is no need to make a mode choice after the trip distribution as is common in the four-step model.

Despite several new approaches in freight flow modelling many freight flow models currently used around the globe are still based on the gravity model or a derivative. One of the input parameters for the gravity model is the decay function. In general this factor is derived from actual data obtained from a statistical bureau or through a census or survey. However the statistical bureau in South Africa does not have the required data and the other methods are currently too expensive and time consuming to be carried out by CSCM or a related institution. The way of deriving decay functions used today by

CSCM therefore lacks scientific background and they are looking for an alternative method.
In the next chapter we will look at the methodology to come to an alternative method for deriving decay functions.

# 4. The methodology

Based on the approaches described in literature and the status quo of the commodity freight flow model it is possible to formulate the approach to answer the central research question in more detail. In Section 4.1 we will first discuss the order of steps in our approach before we use Section 4.2 – 4.5 to go through the steps in detail. In Section 4.6 finally the verification and validation of the proposed model will be briefly discussed.

## 4.1 Order of steps

The proposed approach consists of the following steps:

1) Establish a list of possible factors influencing decay
2) Establish metrics for the measurement of the influencing factors
3) Score the 64 commodity groupings against the influencing factors
4) Conduct a correlation analysis to indicate what factors are related with decay, how they are related (signs) and how strong the relationship is
5) Conduct a regression analysis to indicate what (combination of) factor(s) has an obvious effect on the decay and how strong the effect is of each factor

The analyses in the last two steps are based on all 64 commodity groupings, their decay parameter values and scores on the influencing factors (see Sections 4.4 and 4.5).
The result of this approach will be a regression model including the factors influencing decay and their (relative) importance.

Now we will go through the steps above in somewhat more detail and explain their interrelationship.
The first step in the way to formulate a regression model for the relation between decay (dependent variable, represented by the decay parameter) and the influencing factors (explanatory variables) is to establish a (complete) list of factors that are expected to influence the value of the decay parameter (see Table 4.1). When the list is in place a metric has to be assigned to every factor to be able to measure/score it. Next step will be to score the 64 commodity groupings against the factors based on the assigned metrics. However we will start scoring against the influencing factors that have easy accessible data (given the applied metrics). The other factors might be included later in the process if the initial factors are not sufficient in explaining decay. Scoring the commodity groupings also includes the choice what scoring system to use; a scale or quantitative numbers. The result at this stage will be a table with scores as shown below.

| | Dependant variable | Explanatory variables | | | | | |
|---|---|---|---|---|---|---|---|
| Commodity grouping | Decay parameter | Value per ton | Scarcity | Supply concentration | Local demand | Homogeneity | Brand |
| COAL | 0.05 | 492 | 0.01 | 0.20 | 0.00 | 1.0 | 0.0 |
| STON | 0.04 | 52 | 0.00 | 0.07 | 0.00 | 0.9 | 0.1 |
| IRON | 2.78 | 730 | 0.00 | 0.92 | 0.00 | 1.0 | 0.0 |
| FOOD | 0.15 | 7722 | 0.09 | 0.09 | 0.57 | 0.6 | 0.8 |
| FUEL | 2.00 | 8000 | 0.17 | 0.48 | 0.03 | 0.9 | 0.4 |

Figure 4.1: Visualization of hypothetical results of the proposed approach

When every commodity grouping has a separate score for the different factors we are able to conduct a correlation analysis to indicate what factors are (significantly) related with decay and how they are related (in sign). From this analysis we can also see how the factors influence each other, which is important for the regression analysis.

| | Decay parameter | Value per ton | Scarcity | Supply concentration | Local demand | Homogeneity | Brand |
|---|---|---|---|---|---|---|---|
| Decay parameter | 1 | ,625* | ,692** | -,008 | -,217 | -,235 | ,186 |
| | | ,017 | ,006 | ,978 | ,456 | ,420 | ,524 |
| Value per ton | ,625* | 1 | ,526 | ,270 | ,016 | -,269 | ,524 |
| | ,017 | | ,053 | ,350 | ,956 | ,351 | ,055 |
| Scarcity | ,692** | ,526 | 1 | ,109 | -,249 | -,011 | ,265 |
| | ,006 | ,053 | | ,709 | ,390 | ,971 | ,359 |
| Supply concentration | -,008 | ,270 | ,109 | 1 | -,031 | ,438 | -,146 |
| | ,978 | ,350 | ,709 | | ,915 | ,117 | ,620 |
| Local demand | -,217 | ,016 | -,249 | -,031 | 1 | -,356 | ,494 |
| | ,456 | ,956 | ,390 | ,915 | | ,211 | ,073 |
| Homogeneity | -,235 | -,269 | -,011 | ,438 | -,356 | 1 | -,532 |
| | ,420 | ,351 | ,971 | ,117 | ,211 | | ,050 |
| Brand | ,186 | ,524 | ,265 | -,146 | ,494 | -,532 | 1 |
| | ,524 | ,055 | ,359 | ,620 | ,073 | ,050 | |

Table 4.1: Visualization of hypothetical result of the correlation analysis (step 4)*

* All correlation coefficient tables in this report are derived from SPSS. All commas in these tables represent dots.

The final step of our approach to come to a formula for the decay parameter is to apply a regression analysis to derive a regression model. This model contains the factors that have an obvious effect on the decay parameter and shows the relative weight of this effect on the explanation of the variability in the decay parameter for each factor.

## 4.2 Step 1 and 2 - List of influencing factors and their metrics

Based on literature research and an initial workshop with the persons responsible (within CSCM) for the updates to the model and the decay functions in the past few years (the panel) a list is established of factors possibly influencing the decay. This list of factors has been discussed with people from the development department of Flowmap and colleagues from the logistical department of the university.

Below the factors and possible metrics for their measurement are described. A hypothesis is added to describe the expected relationship between the influencing factor and the decay. We also discuss the independency of the influencing factors as relationships among the influencing factors can disturb the regression analysis.

The goal is to score all commodities on these influencing factors. These scores will partly be based on available data and partly on consensus within the panel (see Section 4.3).

*Metrics*

As briefly mentioned in the previous section a proper metric should be in place to be able to measure variables (in this case referred to as influencing factors). The metric plays a key role in the preparation phase of correlation and regression analyses as the way a variable is measured influences the outcome of the analyses significantly. Because there are no strict guidelines for choosing a proper metric to represent a variable, as it also depends on the available data and the purpose of the measurement, it is a complex step in the approach. It is very well possible that the metrics for the influencing factors need adjustments during the conduction of the study.

***Value per ton***

As mentioned in the literature review (Section 2.3) several studies in the past have found a relation between the value of goods per unit weight and the distance over which they are transported. The assumption is that relative cost of transport decreases when the average value of the product increases and the distance over which the product needs to be transported will have less influence on the choice of the supply point.

The term 'value per unit weight' however can have several interpretations. We could for instance look at the value of production or the value at the moment of consumption. And when a significant part of the total supply of the good is imported we could also choose to work with the value of the goods when they enter the seaport. Moreover we have to choose whether we take all goods from within a commodity grouping into account or only a dominant (representative) product group. Besides that, one could debate about the unit weight. We could for instance take the weight with or without packaging.

None of the past studies put much effort in explaining their precise metrics, leaving us with the choices just described.

Although we seem to have a lot of choice, the available data will limit our possibilities. The data obtained from Conningarth represent the value of products at final production (excluding transportation and packaging). For imported products the value at the moment of landing is used. To obtain these figures Conningarth consults various published and non-published sources. The weight is based on the goods without packaging material.

Therefore we use the production value (combined with import values) of a total commodity grouping divided by the total amount of tons produced (and imported). This is expressed in Rand per ton to be able to make fair comparisons between the commodity groupings.

The value of goods might, for several reasons (like increasing raw material prices), differ over the years. However as we expect the value of goods to influence decay we also expect a change in the decay functions. It is hard to predict if the relation will stay exactly the same over the years. A separate study needs to be carried out to answer this question.

*Hypothesis:*

The higher the average value per ton of the commodity group the lower the value of the decay parameter.

***Scarcity***

If products are not broadly available, consumers (mostly companies in this case) start making long term contracts to have security of supply. These contracts are not necessarily made with the nearest supplier. To increase their supply security even more they tend to spread their risk by contracting different suppliers. In some cases a large part of the supply will even come from overseas via seaports, which are normally not centrally located. That increases the distance of the shipments on road.

There are different ways to divine and measure scarcity. Actual demand (fulfilled + unfulfilled) versus total supply (production + import) could be used as an indicator of scarcity. If the actual demand is much higher than the total supply, scarcity is considered high. However finding out what the actual demand is seems quite a complex exercise.

In some cases the relative price of goods could be used as an indicator for scarcity as products that are scarce are normally more expensive. For example in a desert region fruit will be more expensive (as it is scarce) than at the fruit plantation itself. However detailed data, regarding product prices, need to be available for this type of measuring.

A metric that would be possible to use based on the available data indicates what percentage of the total demand is fulfilled by import. As one could assume that import generally only exists when local production is not sufficient to fulfil total demand.

The scarcity is therefore measured by dividing the tons imported into South Africa by the total demand within the country (including export) per commodity grouping.
One could argue about the relation between scarcity and the value per ton especially in a development country. Cheaper products like raw materials and agricultural goods are mostly produced in the country itself while the more expensive products like electronic equipment and high-tech machinery are often imported. This possible relationship is taken into account during the correlation analysis.

*Hypothesis:*
The higher the scarcity of the products in the commodity grouping is, the lower will the decay parameter be.

***Supply concentration***
The study done by Black (1972) indicated concentration of supply as a factor influencing decay. In case there is only one regional supplier for a certain commodity grouping and the goods produced by this supplier are demanded by all the other regions, the shipments of these goods will have to cover large distances.
We will briefly recall the example given in Section 2.3.
If a region (for example the south-east of South-Africa) is highly specialised in the production of a particular good (for example avocados), it is less likely to be sensitive to distance in its shipments as the avocados have to be transported to the west coast market anyway. A commodity grouping such as stone that is widely available will be dominated by short-haul shipments because of the proximity of competitors.
Black measured 'Supply concentration' by the proportion of flows from the largest producing region (supply in a certain MD) relative to the total amount of flows (total national supply).
If production would be used instead of supply, imports would not be taken into account, which would for some commodity groupings have a significant impact on the flow pattern. From Black's report it does not become clear how the production regions are defined. In his model he works with nine main regions but it is unclear if a production region is equal to one of the nine main regions. We start our study with supply concentration on a MD level.
'Supply concentration' and 'Scarcity' could be related as a high supply concentration could indicate that goods are not everywhere readily available which might suggest a high level of 'Scarcity'. However if a single supply point (no sea port) produces enough to fulfil the total demand of South Africa 'Scarcity' will be (close to) zero. The correlation analysis will show whether a relation is present.
Import figures of products will, like their value, change over time. This does not mean that the relation with decay will change. However it gives rise to additional research for both the value of the decay parameter as well as the import figures over time.

*Hypothesis:*
The higher the concentration of supply, the lower the decay parameter.

***Local demand***
Black (1972) suggested that in case the entire production of a certain good is consumed locally, due to for example economies of scale or perishability, the average distance over which the goods are transported would be low.
For the measurement of local demand Black proposed to use the proportion of the production that is consumed in the same region (MD) as where it is produced; the

intraregional transportation. In this case the same issue as just mentioned for supply concentration arises regarding the size of the production (or local) region. The impact for 'Local demand' is expected to be even bigger. However unlike we can with 'Supply concentration', it is not possible for 'Local demand' to aggregate the MD level into province level. This is due to a lack of data, as we would need to know the flows from all the MD's within a province to each other. Therefore we will use scores based on MD level.

In case a region has a high supply concentration of goods of which the processing (facility) is capital intensive (for example wood) and transportation is relative expensive, one can expect that the processing facility will be built in the region of supply. In other words for some high 'Supply concentration' commodity groupings a high 'Local demand' can be expected. However for processed products that deliver to final customers this relation does not make much sense.

Production facilities can change location however this is expected not to be as variable as the earlier mentioned import figures for example.

*Hypothesis:*

The higher the proportion of local demand, the higher the decay parameter.

***Homogeneity***

The level of homogeneity within commodity groupings is expected to influence the freight flows. Customers will not be willing to pay more for products from further supply points if the local products in their opinion are highly similar.

We look at the homogeneity between sub-commodity groupings from an end-user perspective; to what extend can products within a commodity grouping be used as substitutes for the same purpose. Although there is a difference in grade of wheat between different wheat producing areas, it can all be used for the same purpose. On the other hand boats (part of the commodity grouping of transport equipment) cannot serve the same purpose as a motorbike (same commodity grouping) if we want to travel between home and work. Therefore the commodity grouping wheat will be assigned a higher level of homogeneity than the commodity grouping of transport equipment.

As it is difficult to clearly define homogeneity, scoring this factor is a challenging task. No studies that fit this specific aggregation of commodity grouping have been carried out and subjectivism in the scoring will always be included. We have therefore chosen (also due to time and budget restrictions) to score the commodity groupings based on consensus within CSCM and discuss this with colleagues from the economics and marketing faculty within the Stellenbosch University.

Instead of the end-user perspective it would also be an option to look at homogeneity from a sales location perspective. The homogeneity within a commodity grouping would be high if all sub-commodities would be sold via the same distribution channel for example. Question is if this metric would give enough information regarding the distance over which the goods are transported.

In general high value products are considered more heterogeneous (people put more effort in comparing different offers). However beverages and processed food, both not expensive, are considered highly heterogeneous. While certain minerals are very expensive but highly homogeneous. A very strong relation between 'Value per ton' and 'Homogeneity' is therefore not expected.

'Homogeneity', if scored correctly, would be a very stable factor as the purpose of raw materials and processed products does not change very rapidly.

*Hypothesis:*
The higher the homogeneity within a commodity group, the higher the decay parameter.

***Brand***
Brands are expected to influence the willingness of consumers to (let companies) transport the products they want over further distances (and pay for that). In other words if brands are able to distinguish themselves, customers get brand preferences and finally become brand loyal. In this stage they are willing to consider supply points further away to eventually get their preferred brand.
Like with 'Homogeneity' it is difficult to define a simple but complete metric for 'Brand'. Next to the amount of brands one can also think of brand loyalty, brand awareness and amount of advertisement (in budget for example) among others.
The most important limitation in the choice of the metric is the available data. In our case there are no data readily available within CSCM. We will make use of a similar panel as mentioned above due to similar reasons mentioned for 'Homogeneity'.
"Brand" will be measured by (the combination of) two metrics. We will take an estimation of the amount of brands within a commodity grouping into account as well as a rate for the brand loyalty that is associated with the commodity grouping. Brand loyalty will give additional information about the customers buying behaviour necessary to get a better understanding of brands influence.
These two metrics have been chosen, as they are expected to contain the necessary information regarding the influence of brands on the freight flows and they are most appealing for the members of the panel and therefore easier to score.
In general 'Homogeneity' and 'Brand' are expected to be related as more brands, types and variations indicate more heterogeneity. However as we look at 'Homogeneity' from an end-user perspective this relationship might not be very apparent.
'Brand' and 'Value per ton' might show a relationship as products of high value (for example motor vehicles) often have higher brand loyalty. However the amount of brands in processed food and beverages are very high but the value per ton is rather low.
The amount of brands as well as the brand loyalty per commodity grouping can be considered stable. Therefore also 'Brand' as possible influencing factor is expected not to change much over time.

*Hypothesis:*
The more brands are present in a commodity group in combination with a high brand loyalty the lower the decay parameter.

In some cases characteristics of sub-commodities within one commodity grouping can differ a lot. An example is the difference between a bicycle and a boat, both part of the commodity grouping transportation equipment. These different characteristics might influence the scores of the possible influencing factors (for example the value per ton). In that case an indication of the inequality of the distribution (of value per ton over the sub-commodities) might be useful. This could for example show that only one sub-commodity increases the value per ton for a whole commodity grouping significantly. A method used for this purpose is known as the Gini coefficient (see Appendix C for more detail). However a lot of in-dept industry knowledge is necessary to be able to use this coefficient in an appropriate way. Because of time limitations we will therefore have to leave this additional instrument out of the research for now.

Now that we have defined a separate metric for each of the influencing factors we are able to score the different commodity groupings against the influencing factors.

### 4.3 Step 3 - Scoring the commodity groupings

The process of scoring the commodity groupings depends mainly on the way the metrics, to measure the factors, are being formulated. Scores for 'Value per ton', 'Scarcity', 'Supply concentration' and 'Local demand' can be derived directly from the available data (hard, quantitative data) at CSCM. However as mentioned earlier scores for 'Homogeneity' and 'Brand' will be established within CSCM by organizing a panel discussion (on scales, see Appendix D). Although the scientific value of this method is limited it might give a reasonable good indication whether further research on these two factors will be useful.

The end result of the scoring procedure may have the form of the table below.

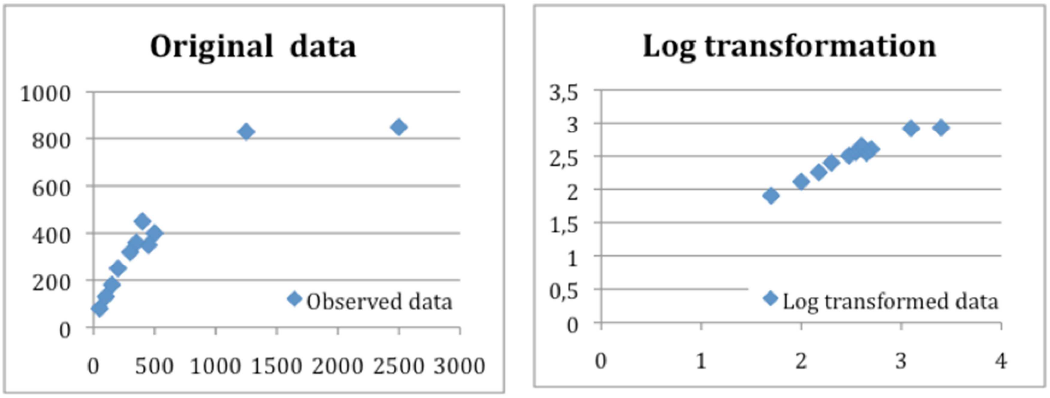
| Commodity grouping | Decay parameter | Value per ton | Scarcity | Supply concentration | Local demand | Homogeneity | Brand |
|---|---|---|---|---|---|---|---|
| BARLEY | 0.5 | 2416 | 0.35 | 0.54 | 0.03 | 0.9 | 0.02 |
| COTTON | 0.15 | 4000 | 0.31 | 0.55 | 0.18 | 0.9 | 0.02 |
| DECIDUOUS FRUIT | 0.15 | 4243 | 0.0 | 0.4 | 0.12 | 0.75 | 0.1 |
| CITRUS | 0.15 | 2832 | 0.0 | 0.32 | 0.0 | 0.8 | 0.1 |
| SUBTROPICAL FRUIT | 0.15 | 4620 | 0.02 | 0.1 | 0.0 | 0.8 | 0.1 |
| VITICULTURE | 0.15 | 3438 | 0.0 | 0.14 | 0.0 | 0.9 | 0.05 |
| GRAIN SORGHUM | 0.50 | 1818 | 0.14 | 0.14 | 0.0 | 0.9 | 0.02 |
| LIVESTOCK | 0.01 | 20870 | 0.1 | 0.05 | 0.06 | 0.7 | 0.15 |
| MAIZE | 4.0 | 2500 | 0.0 | 0.1 | 0.17 | 0.9 | 0.05 |
| SOYA BEANS | 0.5 | 4023 | 0.19 | 0.15 | 0.01 | 0.9 | 0.02 |
| SUNFLOWER SEED | 0.5 | 4272 | 0.09 | 0.06 | 0.0 | 0.9 | 0.02 |
| VEGETABLES | 0.01 | 2200 | 0.02 | 0.06 | 0.0 | 0.7 | 0.15 |
| WHEAT | 0.05 | 4000 | 0.38 | 0.15 | 0.0 | 0.85 | 0.02 |
| ... | ... | ... | ... | ... | ... | ... | ... |

Table 4.2: Scoring of commodity groupings against possible influencing factors

In case scoring is done with hard numbers (quantitative data) there will probably be outliers in the data. Next we explain how we will deal with them.

*Outliers and transformations*

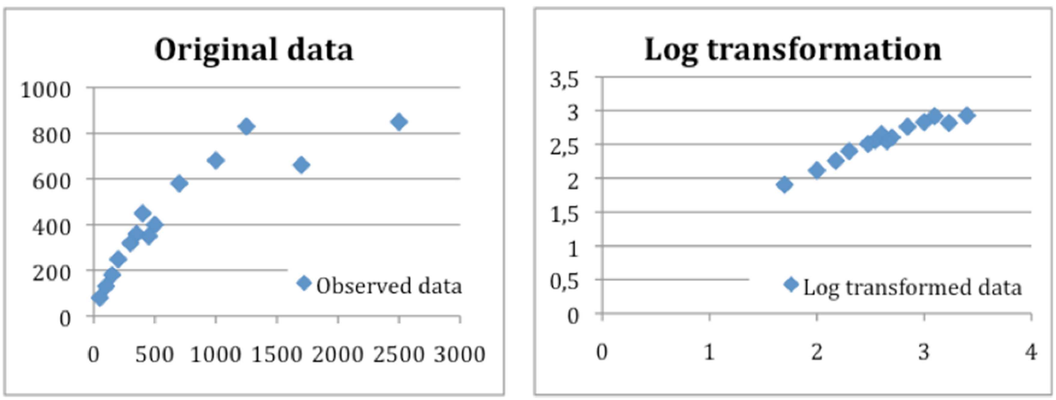
Because extreme values (outliers) can seriously influence the outcome of the next step, the correlation analysis, it is important to give them sufficient attention. An often-used method to deal with extreme values is to use one-to-one transformations. A commonly applied transformation in the steps towards a correlation analysis is the log-transformation. This transformation overcomes the problem of extreme ranges between values of interest. It enables the user to spread out the cluster of low values and pulling in the higher values (Taaffe, 1996). The transformation results in the regression equation $\log Y = a + b \log X$, which becomes $Y = A^*(X)^b$ when taken out of logs. In some cases, depending on the data, this transformation provides a better description of the relation between Y and X than did the original untransformed equation of $Y = a + b^*X$ as shown in the fictional schematic example graphs below.




(a) (b)

Figure 4.2: Visualisation of log transformation of linear relationship

An even more common reason to use one-to-one transformations appears when the observed relationship is obviously curvilinear rather than linear.




(a) (b)

Figure 4.3: Visualisation of log transformation of curvilinear relationship

Now that we have scored all the commodities against the influencing factors and transformed the data in such a way that the relationship between the influencing factors and the decay parameter is best described we can verify the hypothesis concerning the relations. Therefore a correlation analysis will be conducted.

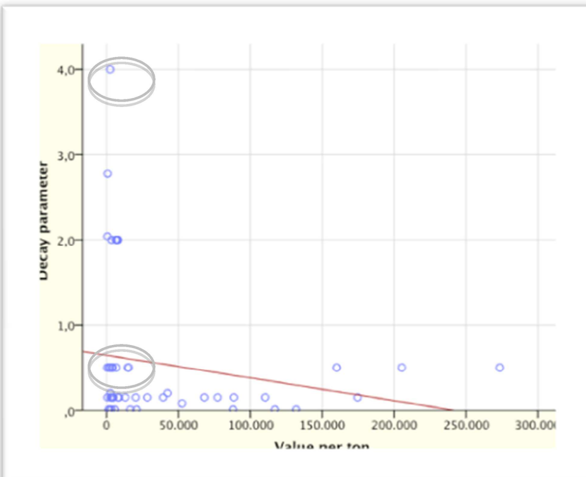
## 4.4 Step 4 - Correlation analysis

A correlation analysis is meant to measure the linear relationship between two variables, resulting in a correlation coefficient (Gujarati, 2006):

(4.1) $\rho = \mathrm{cov}(X, Y)/\sigma_x\sigma_y$

The coefficient has a value between -1 and 1, the former meaning perfectly negative correlation and the latter meaning perfectly positive correlation. If two variables are (statistically) independent the coefficient will be zero. However the opposite is not true as two variables can have a relation that is non-linear (and therefore has a correlation coefficient of 0), for example $Y = X^2$.

We will start by analysing the relation between the decay parameter and the influencing factors one by one in a so-called univariate analysis. This means we have to conduct six separate correlation analyses. To visualize these correlations we will create scatter plots, commonly applied for this purpose. In the example below we visualize a correlation analysis for the relation between ‘Value per ton’ and the values of the currently used decay parameters. All commodity groupings that have initially been assigned a negative power decay function have been included. Therefore every observation (blue dot) in the scatter plot represents a certain commodity grouping.




**Correlations**

| | Value per ton |
|---|---|
| Decay parameter | -,193 |
| | ,084 |

Figure 4.4: Fictive relationship between value per ton and decay (left)*

Table 4.3: Correlation coefficient table (right)

* All scatter plots in this report are derived from SPSS. All commas in these figures represent dots.

From the scatter plot (Figure 4.4) we can see a very weak relation between the value per ton of a commodity grouping and their assigned decay parameter values. The graph shows a remarkable pattern; commodity groupings of similar value per ton have all been assigned different decay parameters. Compare for example the commodity groupings barley ($\beta$ = 0.5) and maize ($\beta$ = 4.0), both having a value of circa R2500 (see circles in Figure 4.4 and values in Table 4.3).
In this section the scatter plot is only used as a visualization example. In the next chapter we will look into this matter in more detail.
The correlation coefficient presented in Table 4.3 confirms our observation from the scatter plot. The value of -0.193 indicates a very weak negative relation between the value per ton and the value of the decay parameter. The significance (0.084) is not important for now and will be explained in more detail in Chapter 6.

For the initial correlation analysis the currently applied decay parameters will be used. As a rule of thumb used by statisticians the minimum amount of observations needed for a correlation analysis is ten per variable (influencing factor) studied. In total 64 commodity groupings are available, meaning 64 observations. However 11 of these commodity groupings have been assigned a negative exponential decay function and 53 have been assigned a negative power decay functions. As the decay parameters of the two functions in their current format are not comparable we will have to split the correlation analysis; one group for each type of assigned decay functions.

For a univariate analysis both observation groups are of sufficient size. However in the regression analysis (multivariate analysis) we might include several factors (that were of significant influence in the univariate analysis) at the same time. For this type of analysis the group with negative exponential decay functions will be too small.
In case more than one factor seems to significantly influence decay, we can choose to use only the group with negative power decay functions or we have to assign a negative power decay function to the commodity groupings that currently have been assigned a negative exponential decay function. The functions will not be the same (having a different form/shape) but it will be the closest approximation.
In both cases the total amount of observations would be sufficient to handle the inclusion of at least five factors in a multivariate analysis may that be the outcome of the univariate analysis.
A standard significance level of 0.05 will be used for in or exclusion of factors for the multivariate analysis.

If a factor is highly significant in the univariate analysis but is not significant in the regression analysis, in combination with other factors, this normally indicates a high degree of multicollinearity between the factors. A high degree of multicollinearity means that there is a high correlation between two or more explanatory factors. A fictive example is given for ‘Scarcity’ and ‘Value per ton’ in Table 4.4 and Table 4.5.

| Variables | Value | Scarcity | Supply concentrat ion | Local demand |
|---|---|---|---|---|
| Decay | 0.959105 | 0.906870 | -0.565907 | -0.381906 |

| Variables | Value | Scarcity | Supply concentrat ion | Local demand |
|---|---|---|---|---|
| Decay | 0.927309 | 0.014248 | -0.020167 | -0.021743 |

Table 4.4: Correlation coefficient table after the univariate analysis (left)

Table 4.5: Correlation coefficient table after the regression analysis (right)

We can see from Tables 4.4 and 4.5 that the relation between decay and ‘Scarcity’ was strong during the univariate analysis while the relation was low in the regression analysis, indicating multicollinearity. To overcome this problem we will have to carry out extra univariate analyses between the explanatory factors. This can either be done on beforehand or when the regression analysis indicates multicollinearity.

The table below shows the correlation among all influencing factors. ‘Scarcity’ and ‘Value per ton’ are correlated but the relationship is not very strong.

| | Decay parameter | Value per ton | Scarcity | Supply concentration | Local demand | Homogeneity | Brand |
|---|---|---|---|---|---|---|---|
| Decay parameter | 1 | ,625* | ,692** | -,008 | -,217 | -,235 | ,186 |
| | | ,017 | ,006 | ,978 | ,456 | ,420 | ,524 |
| Value per ton | ,625* | 1 | ,526 | ,270 | ,016 | -,269 | ,524 |
| | ,017 | | ,053 | ,350 | ,956 | ,351 | ,055 |
| Scarcity | ,692** | ,526 | 1 | ,109 | -,249 | -,011 | ,265 |
| | ,006 | ,053 | | ,709 | ,390 | ,971 | ,359 |
| Supply concentration | -,008 | ,270 | ,109 | 1 | -,031 | ,438 | -,146 |
| | ,978 | ,350 | ,709 | | ,915 | ,117 | ,620 |
| Local demand | -,217 | ,016 | -,249 | -,031 | 1 | -,356 | ,494 |
| | ,456 | ,956 | ,390 | ,915 | | ,211 | ,073 |
| Homogeneity | -,235 | -,269 | -,011 | ,438 | -,356 | 1 | -,532 |
| | ,420 | ,351 | ,971 | ,117 | ,211 | | ,050 |
| Brand | ,186 | ,524 | ,265 | -,146 | ,494 | -,532 | 1 |
| | ,524 | ,055 | ,359 | ,620 | ,073 | ,050 | |

Table 4.6: Correlation coefficient table for univariate analyses among influencing factors

If two variables are both significant in the multivariate analysis there may still be a chance for multicollinearity, but this would tend to be small as the one variable would already contain most of the information. It is important to detect multicollinearity as including linear correlated variables in the final regression model can cause large standard errors in the estimation of the parameters of the influencing factors. Especially as the purpose of the study is not just to forecast a mean value for the decay parameter but to obtain a reliable estimation of the individual parameters included in the model.

It is important to note that correlation does not necessarily imply causality. The expected influencing factors can be correlated with decay without really influencing it. An example is the size of the trucks used for the transport of the commodity groupings; although the size of the truck increases if the transport distance increases (decay parameter becomes lower) because it is economically more interesting, it does not influence the value of the decay parameter itself. If a truck company would decide to use bigger trucks (because it was a special offer) it does not mean that the distances goods are transported suddenly increase. In that case the factor can still enter the model but behaves more like a predictor. The factors in this research are expected to actually influence decay. If for example the supply gets more concentrated (because of the merging facilities) the distance the produced goods are transported is expected to increase. This lowers the value of the decay parameter.

A thorough investigation with an open mind is necessary to make sure that also hidden correlations are brought to the surface. It is for example possible that certain factors do not show a clear correlation when looking at the total set of commodity groupings but when excluding the low value commodity groupings an obvious relation would appear.

### 4.5 Step 5 - Regression analysis

Regression analysis is as mentioned earlier concerned with the study of the relationship between one variable called the dependent variable and one or more other variables called explanatory variables. To be able to indicate a relation between the variables, data from the total population or a sufficient part of it are necessary. If data from the total population are available it is possible to establish a population regression function (PRF) (Gujarati, 2006).

(4.2) $E(Y|X_i) = B_1 + B_2X_i$

$E(Y|X_i)$ is the expected value of Y corresponding to a given value of X. $B_1$ and $B_2$ are called the parameters, also known as the regression coefficients. $B_1$ is referred to as the intercept (coefficient) and $B_2$ as the slope (coefficient).
If we want an expression for the relation on a level of individuals within a population an extra component besides the mean value of the population needs to be included. This component represents the deviation of the individual case from the mean of the population. Therefore the PRF function is now of the following form:

(4.3) $Y_i = B_1 + B_2X_i + u_i$

The extra component $u_i$ is known as the stochastic error term. Besides an error due to the randomness of human behaviour the stochastic error term may represent the influence of

those variables that are not explicitly included in the model. Sometimes these variables are not known but often these variables are left out because their explaining value is little and inclusion would make the model unnecessarily complex.
As an example randomness of human behaviour in explaining freight flows can occur due to a lack of information; customers are not aware of the closest supply point.

In case it is not possible to base the regression on the total population a sample data set is used. As just a sample of the total population is known it is only possible to make an estimation of the population parameters. This concept is known as sample regression function (SRF). The stochastic version of the equation may be written as (Gujarati, 2006):

(4.4) $Y_i = b_1 + b_2X_i + e_i$

Where $e_i$ represents the difference between the actual Y values and their estimated values from the sample regression ($e_i = Y_i - \hat{Y}_i$, where $\hat{Y}_i = b_1 + b_2X_i$). The parameters $b_1$ and $b_2$ are the estimators of respectively $B_1$ and $B_2$.

Extra explanatory variables can easily be added to the expressions stated above turning the simple or two-variable regression models into multiple linear regression models. The stochastic sample regression function may for example have the following form:

(4.5) $Y_i = b_1 + b_2X_{2i} + b_3X_{3i} + b_4X_{4i} + b_5X_{5i} + e_i$

Linear regression can be interpreted in two ways; linearity in variables or linearity in parameters. The PRF and SRF expressions above are linear in both variables and parameters. The expressions below show examples of non-linearity in variables (first) and non-linearity in parameters (second).

(4.6) $Y_i = B_1 + B_2X_i^2 + u_i$
(4.7) $Y_i = B_1 + B_2^2X_i + u_i$

The economic theory is usually not strong enough to tell us the functional form in which the dependent and explanatory variables are related. By choosing the wrong functional form, the estimated coefficients may become biased estimates of the true coefficients (Gujarati, 2006).
We will start with an expression for the decay parameter that is at least linear in his parameters. If we obtain indications for the use of another model expression this form will explicitly be mentioned and tested. In that case variables do not necessarily have to enter the model linearly.

*Dummy variables*
The regression model can consist of quantitative or numerical as well as qualitative variables. Qualitative variables (e.g., gender, colour, nationality) are represented by dummy variables. These variables can take on values of 0 or 1, indicating the absence or presence of that attribute. For the rest they behave similar to the numerical variables; they have a weight factor and can have a non-linear relation with the depending variable. The dummy variables are not applicable for the current possible influencing factors but other factors in a later stage might come in such format.

*Ordinary least square method*

To estimate the parameters $b_1$ and $b_2$ of the SRF expression stated above different methods can be applied. One of the most commonly used methods is the ordinary least square method (OLS). This method estimates $b_1$ and $b_2$ in such a way that the residual sum of squares, $\Sigma e_i^2$, is as small as possible (Gujarati, 2006). The principle states:

(4.8) Minimize $\Sigma e_i^2 = \Sigma(Y_i - b_1 - b_2 X_i)^2$

When the mutually uncorrelated influencing factors have been identified and the regression analysis including the application of the OLS method resulted in a regression model a goodness-of-fit test needs to be carried out to ensure that the regression model is sufficiently accurate in describing the decay parameter.

The measure developed for this purpose is known as the coefficient of determination denoted by the symbol $R^2$ (R-squared). $R^2$ measures the proportion or percentage of the total variation in Y explained by the regression model. In other words, it indicates the degree to which the explanatory variables explain the value of the decay parameter. The equation for $R^2$ is:

(4.9) $R^2 = 1- (\Sigma(y_i - \hat{y}_i)^2 / \Sigma(y_i - \bar{y})^2)$ ( = 1- (residual sum of squares / total sum of squares))

Where $\Sigma(y_i - \hat{y}_i)^2$ is the variability about the regression line that is left over after the regression model is fitted. $\Sigma(y_i - \bar{y})^2$ is the total amount of variability in the response variable Y (in other words the sum, over all observations, of the differences of each observation from the overall mean).

This equation can be rewritten into:

(4.10) $R^2 = (1 - \Sigma e_i^2)/ \Sigma y_i^2$

As we know the residual sum of squares from the earlier applied OLS, $R^2$ can now be computed.

Most of the previous steps can easily be performed using a statistical program (like SPSS). Nevertheless it is good to know the theory behind the presented output.

In case the factors included in the regression model, based on their significant correlation with the decay parameter, do not explain enough about the decay this could be an indication that the initial list of potential influencing factors was not complete or that the metrics used for the measurement of the factors were not chosen properly. It could also be possible that the factors are right but the relation between the depending and explanatory factors is non-linear (see Figure 4.4 below), in that case the functional form of the model should be adapted (which is a complicated task).

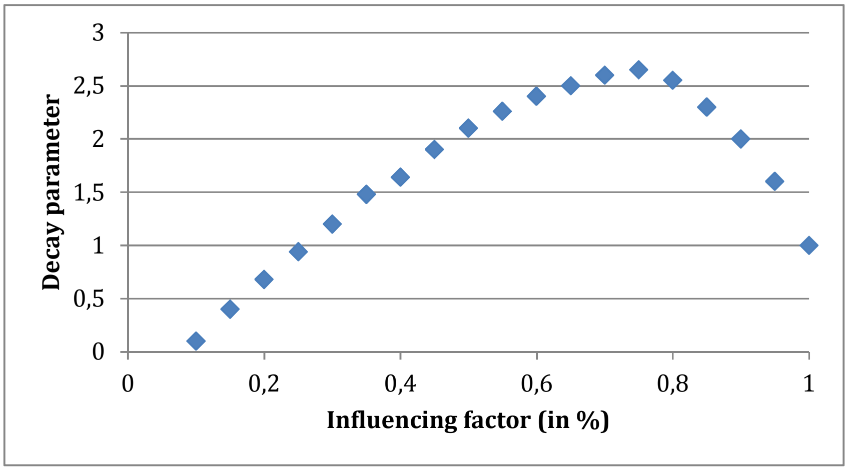


Figure 4.5: Nonlinear relation between decay and a possible influencing factor

Besides choosing the wrong functional form errors in the model can obviously arise by errors of measurement in the dependent or explanatory variables. Errors in the measurement of the dependent variable will increase the estimated variances of the OLS estimators because an extra error gets added to the common error term, $u_i$. However in practice this error does not seem to matter much (Gujarati, 2006).

Errors in the measurement of explanatory variables make the OLS estimators biased and inconsistent. That means a more serious problem. And as no real remedies exist, making sure that the gathered data for the explanatory variable are as accurately as possible is of the utmost importance.

## 4.6 Verification and validation

If the regression model passes the goodness-of-fit test it is sufficiently accurate in describing the decay parameter from a statistical point of view. However carrying out a reality check on the formula is a critical next step. The reality check should include a test of robustness:

* How does the decay parameter (and therefore the average distance goods are transported) change when the value of an influencing factor for a certain commodity grouping increases or decreases with a certain percentage.

Verification and validation will be discussed in Section 6.3.

## 4.7 Chapter summary

The proposed alternative method for deriving decay functions is based on regression modelling. To come to a model that is able to construct decay functions based on the characteristics of commodity groupings the possible influencing factors must be listed, metrics for the measurement of the influencing factors need to be established, the commodity groupings need to be scored against the influencing factors and correlation and regression analyses must be carried out. As the dependent variable for the regression model the decay parameter is chosen. From the factors expected to influence the decay parameter, the explanatory variables, ‘Value per ton’, ‘Scarcity’, ‘Supply concentration’ and ‘Local demand’, are based on quantitative data while ‘Homogeneity’ and ‘Brand’ are scale values and based on the outcome of a panel discussion.

When the regression model is established it needs to be extensively tested before the outcomes can be used in the commodity freight flow model.

The intermediate results and final outcome of the described methodology are presented in Chapter 6. However in the next chapter we will first discuss the data used as input for the different steps of the method.

# 5. Data

In the previous chapter we have explained the methodology that should lead to a regression model for the decay parameters in the South African commodity freight flow model. To be able to score the commodity groupings against the influencing factors (step 3) and conduct the described correlation analysis (step 4), specific input data are required. In this chapter we will list the necessary data and their sources (Section 5.1) and try to verify their quality (Section 5.2). The present input data that do not fulfil the expected quality are compared with figures used in similar models (Section 5.3). In Section 5.4 ways to obtain alternative data are proposed. The results of the data gathering and the data processing are presented in Section 5.5 and 5.6 and the currently used and newly obtained decay functions are compared in Section 5.7. This chapter will be concluded by a robustness analysis of the newly obtained decay functions (also in Section 5.7).

## 5.1 Required data

The proposed metrics in step 2 of the presented methodology implies the availability or gathering of commodity grouping related data. We discuss the required data for the metrics below followed by the additional data required for conducting the correlation analysis; the decay functions.

### *5.1.1 Metrics*

The required data for every proposed metric differs but their main source is similar. All data are initially provided by Conningarth. The data requirements count for all commodity groupings as they all need to be scored.

*Value per ton*

* Average value per ton; derived from the total value of supply of a commodity grouping divided by the total tons supplied

*Scarcity*

* Total imports
* Total demand; one figure readily available (containing total consumption, intermediate demand, total investment (inventory) and total imports)

*Supply concentration*

* Total production per MD

*Local demand*

* Total production per MD
* Total consumption per MD

### *5.1.2 Decay functions*

To be able to analyze the relationship between the influencing factors and decay, a decay parameter is required for every commodity grouping involved in the correlation analysis. As stated in the previous chapter we will initially use the decay parameters applied by CSCM in the commodity freight flow model.

## 5.2 Data quality verification

*Influencing factors*

All the required data for the scoring of the proposed metrics are provided by Conningarth. Conningarth's role has briefly been discussed in Chapter 3. They do not give full insight in their model, as this is the driver for their business. However in case of specific questions regarding certain supply or demand figures they generally reveal detailed figures from reliable sources. Although they have to make assumptions in their econometric model, that might be debatable, they have a dedicated team of more than 40 people with various industry backgrounds full time working on the model. A source that can definitely compete with a nation-wide survey or census when looking at the (disaggregated) supply and demand figures.
The data from Conningarth, as described in the requirements above, have been reviewed and seem appealing.

*Decay functions*
As the lack of scientific background for the use of the current decay functions was one of the main reasons to start this research verification of the quality of these factors was required. As the amount of known freight flows was not sufficient to verify the quality of the decay functions, findings from previously performed scientific studies have been used. These studies mentioned in the literature review suggest and confirm relations between decay parameters and certain influencing factors ('Value per ton', 'Scarcity', 'Supply concentration' and 'Local demand', see also Section 4.3). The same relations were expected to be found using the provided data from Conningarth and the currently applied decay parameters. A visualization and summary of the results of the conducted correlation analyses are presented below. We initially concentrated on the possible influencing factors that could be scored based on readily available data. The correlation analysis between 'Homogeneity', 'Brand' and decay will be discussed in the next chapter.

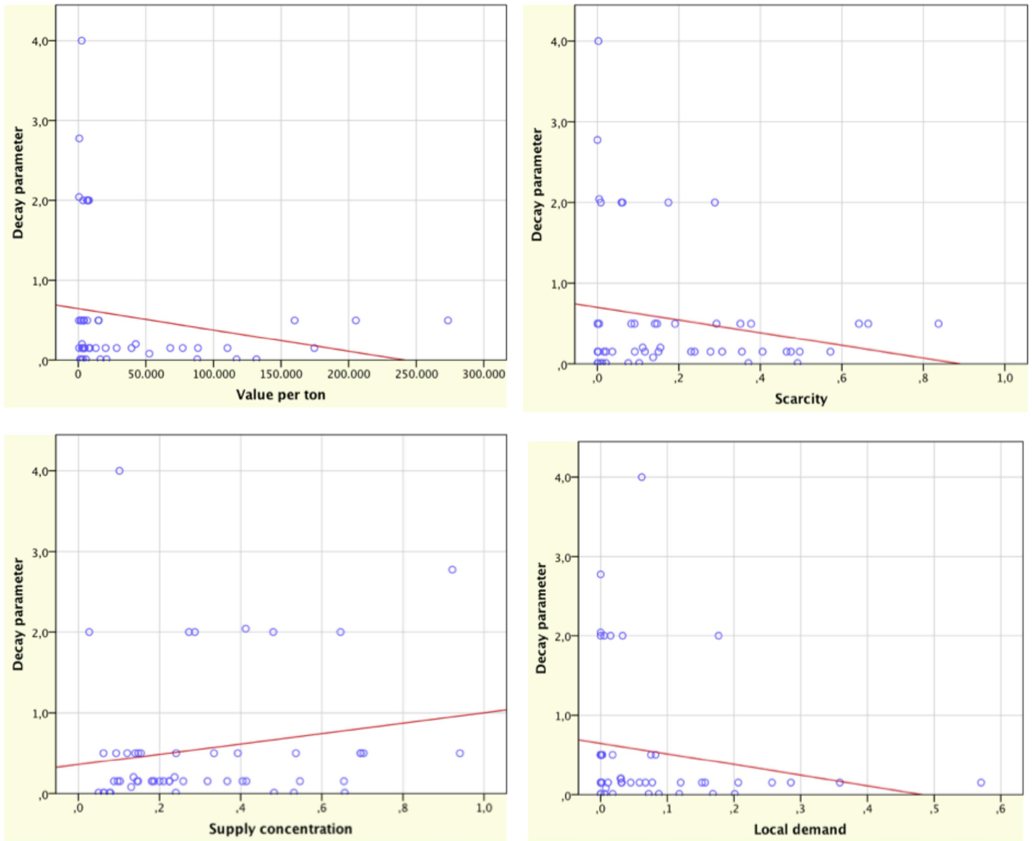


Figure 5.1: Correlation analysis between proposed influencing factors and decay*

| | | Value per ton | Scarcity | Supply concentration | Local demand |
|---|---|---|---|---|---|
| Decay parameter | Pearson Correlation | -,193 | -,198 | ,182 | -,176 |
| | Sig. (1-tailed) | ,084 | ,077 | ,096 | ,104 |
| | N | 53 | 53 | 53 | 53 |

Table 5.1: Correlation coefficient table*

* Based on the decay parameters of the 53 initially applied power functions

The correlation analysis presented above is based on the 53 commodity groupings that were initially assigned a negative power decay function. Using the negative power decay functions gives us a sufficient set of observations and the possibility to compare the decay parameters with other studies (that have applied negative power decay functions, see Section 5.3). In Appendix E the outcome of the correlation analysis based on the 11 negative exponential decay functions can be found. In the same appendix we also present the results of the correlation analysis based on all 64 commodity groupings (where the commodity groupings with a negative exponential functions are assigned the best fitting negative power functions). Both analyes do not result in better correlations.

From the correlation coefficient table (Table 5.1) it becomes clear that none of the influencing factors has a significant relation (given a significance level of 0.05) with the decay parameters currently used in the commodity freight flow model. The scatter plots (Figure 5.1) reveal one of the underlying reasons: commodity groupings with very

different characteristics (or scorings on the possible influencing factors) have been assigned the same decay parameters.
While the commodity groupings with assigned decay parameters higher than 2 (maize, iron ore, chrome, fertilizers, petroleum, gas, non-ferrous metal basic industries and jet fuel) seem to be the outliers, they are (for the first three factors) actually in line with the hypotheses. From that perspective all commodity groupings in the lower left corners of the first three graphs are the ones that are 'far off' (see Appendix A for the related commodity groupings).
There is no clear indicator (common characteristic) among the commodity groupings, which got assigned a high decay parameter that would justify a different treatment from the other commodity groupings at this point.
An explanation for the phenomenon just observed could be that the scoring of the commodity groupings on the influencing factors is incorrect. As mentioned above the data provided by Conningarth seem relatively reliable and appealing. It is however possible that the definitions of for example value per ton and local demand differ between this study and the ones in the past (as none clearly described the definitions used). Instead of the data it is also possible that the differences in methodology between the studies influenced the correlation analysis. The size of the regions, grouping of the commodities and the use of the Euclidian distance may have altered the expected relationships between decay and possible influencing factors. Another explanation could be that the assignment of the decay parameters to the commodity groupings has been based on other characteristics than used in this analysis. From the perspective of this analysis they should therefore be considered as incorrect. A last explanation could be that the relationships between influencing factors and decay as confirmed in earlier studies do not apply for the current situation in South Africa.
One or more of the four possible explanations (inaccurate data, wrong methodology, incorrect assignment of decay parameters and inapplicability to current South Africa) might be the cause of the non-existence of correlation between decay and the possible influencing factors. However due to many (time and data) limitations we will take the data for granted for now as well as the applied methodology (although size of the regions has been researched, see Appendix F). The current opinion within CSCM is that the hypotheses as stated in the previous chapter should apply for South Africa in her current state as well. Therefore we will first focus on a more thorough quality check of the currently applied decay parameters.

## 5.3 Comparison of decay parameters

A first step in the quality check of the currently used decay parameters is to compare them with other decay parameters obtained and used in similar models. In Section 2.4 the models of Black (1972) and De Jong et al. (2010) have been discussed and their related decay parameters have been listed in Appendix B. Publication of applied decay parameters in freight flow modelling is scarce and the comparison we want to use them for is indicative (as differences between the models make a meaningful comparison difficult) therefore we will limit ourselves in this comparison to the decay parameters related to the two models just mentioned.

As the level of aggregation in commodity groupings influences the spread between the decay parameter values (Black, 1972) it makes more sense to compare the relative values (high/low) of the parameters instead of the actual values.

The high values for Fertilizers and Pesticides (#43), Sugar Cane (#16) and Stone Quarrying, Clay and Sand-Pits (#28, #29, #30) are confirmed by the studies of Black while the value for Fertilizers is assigned a low value in the Dutch model. Petroleum Refineries and Products of Petroleum/Coal (#45) are assigned a high level value in the BasGoed model too, however in both studies of Black they have been assigned a low value. 'Bricks' is part of Structural Clay Products (#325) in Blacks' commodity classification and is therefore expected to have a very high parameter value. However in the South African model the parameter value of 'Bricks' is only average (0.60). A clear reason for this deviation cannot be given at this state.

Besides these high values also the low values that are assigned in South Africa to Cotton (#2), Motor vehicles (Parts and Accessories (#58, #59) Machinery and Equipment (#56) and Textiles, Clothing, Leather Products and Footwear (#37) are confirmed by the other studies.
The Dutch model is the only model confirming the low value for Food and Food Processing (#34). CSCM assigned a low value to Pharmaceutical, Detergents and Toiletries (#44). This is in line with Drugs (#283) in the commodity classification of Black. However when paints and other chemicals are added, like in the Shippers classification (#8), the value of the decay parameter becomes very high.

Remarkable differences between the applied decay parameters are the values assigned to Dairy (#15) and Poultry Products (#14). Both are assigned a very low value by the CSCM while they have amongst the highest parameter values in the commodity classification. It is very well possible that keeping cattle was more usual in the US in 1967 than it is in South Africa at the moment.

The low values given to Beverages (#35) and Tobacco (#36) are also in contrast with the outcome of Black's studies. In the Shipper classification they are scored amongst the highest values. It might be that both product groups were produced in more but smaller production facilities 40 years ago in the US than they are now in South Africa (where almost all tobacco products are produced in one single facility).

As we have seen there are some clear similarities and deviations between the applied decay parameters. We have not looked at the possible differences in scores on the influencing factors (as we do not have scores for the other studies) but the differences in the parameters can already significantly influence the correlation between decay and the possible influencing factors. The observed differences in the decay parameters suggest that there are either significant differences between the countries and/or time periods or that the assignment of the decay parameters in South Africa have been incorrect. In the next section we will present an approach to test this last hypothesis.

### 5.4 Obtaining actual freight flow data

As an approach to test if the assignment of the current decay functions has been incorrect we propose to obtain actual freight flows from the industry, derive decay functions from these data and comparing the new obtained decay functions with the ones currently applied.

Obtaining actual freight flows is considered as a difficult task as for most industries this type of information is not centrally recorded and for the individual players in the market there is very little gain in supplying the data so fulfilling the request for information has low priority. Apart from that there are often issues around confidentiality and the bigger the organisation the more this type of data is scattered over different divisions. Moreover to obtain sufficient information on a commodity grouping with many sub-commodities many different players need to be contacted.

It would be preferable to derive decay functions for the commodity groupings of which the decay parameters showed the biggest deviation from the parameters obtained in the other studies (see previous section). However given the time restrictions of this project we will limit the data gathering process to the commodity groupings of which we expect to have the highest potential of obtaining sufficient data. Therefore we pick the commodity groupings:

- for which we have relevant contact persons
- with industry boards
- with dominating companies in the market
- with limited supply and/or demand points

The commodity groupings we have focused on and a description of their industry are listed in Appendix G.

The data we try to obtain are preferable detailed actual origin-destination data including distances and volumes transported. Although distances and volumes transported are sufficient to derive decay functions, the actual origin and destination data can give better insights in the actual movements of the commodity groupings and therefore help in understanding the industry characteristics. In addition these figures are more valuable for the reality check of CSCM on the modelled flows. The data should preferably cover the whole range of sub-commodities otherwise at least the majority. We consider the obtained freight flow data sufficient if the accumulated data for a certain commodity grouping represent at least 30% of the total tons transported of that commodity grouping within South Africa. Census and surveys generally accept even a lower percentage.

### 5.5 Processing obtained data

Before the obtained data can be used for the derivation of the decay functions they need to be processed in the right format. As the data come from a variety of boards, associations, research institutes, governmental bodies and commercial organizations the way it is provided differs significantly. Some of the provided data existed out of three columns: origin, destination and volume. Based on MapSource (GIS software) and a shortest routes criterion the distance between the origin and destination was calculated. But in several cases confidentiality limited the companies in giving exact points of origin and destination. In some of those cases only distances and tons transported were provided or in cases of high level locations (provinces) an average distance needed to be derived, normally using the middle points of the indicated regions. Depending on the industry the volumes were not in tons but in amount of units. In those cases average weights had to be obtained to be able to convert the amount of units into tons.
In Appendix G assumptions are listed for every set of obtained data to be able to use the information for the next step; deriving decay functions.

After the distances and transported volumes of the different sources for a certain commodity were converted into the desired format they were combined in a volume vs.

distance graph as shown below. These graphs form the input for the derivation of decay functions.

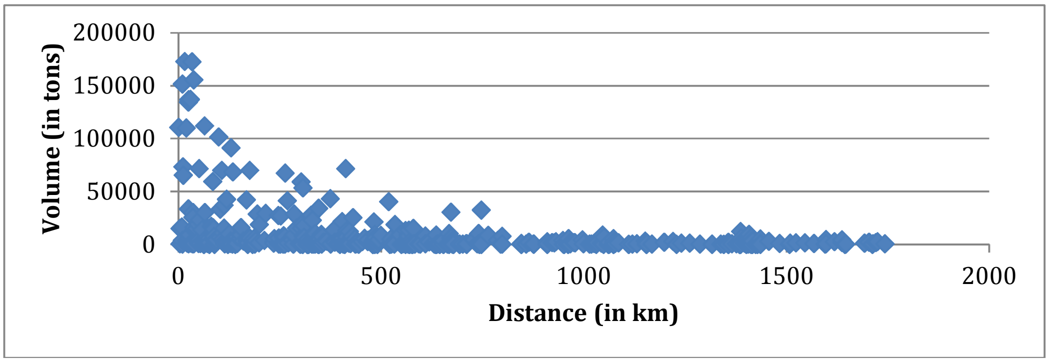



Figure 5.2: Unclassified observations for beverages

As mentioned before the decay parameter generally indicates how transported volume decreases over distance. Figure 5.2 seems to be in line with this definition. To derive a decay parameter from this graph we need to fit a trend line (or decay function) to the observations.

However we cannot use this graph for the derivation of the decay parameters yet. This has to do with the fact that all the observations are equal-weighted. If a line (a negative power function for example) would be fitted to these observations, a cluster of many low volume-short distance observations and a couple of high volume-short distance observations, it would go through the low volume cluster (as shown in the figure below)

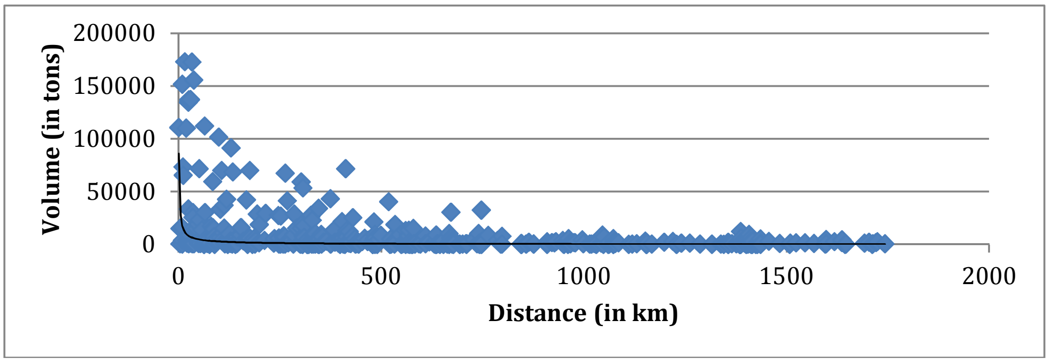



Figure 5.3: Fitting curve to unclassified observations for beverages

This contradicts with our definition of decay and the visual presentation in Figure 5.2. The remedy is to add up the low volume-short distance observations with the high volume-short distance observations before we fit a trend line to the data.

In this procedure small distance buckets are defined (for example buckets of 20km) and the volume of all observations within each bucket is added-up. The result of this procedure, called binning, is illustrated by the graph below.

The size of the bins depends on the provided data and the distance over which the goods are being transported. The more detailed the data and smaller the spread, the smaller the bin sizes. In Appendix H the commodity groupings beverages and deciduous fruit are used as example to show the influence of the binning procedure on the decay parameter.

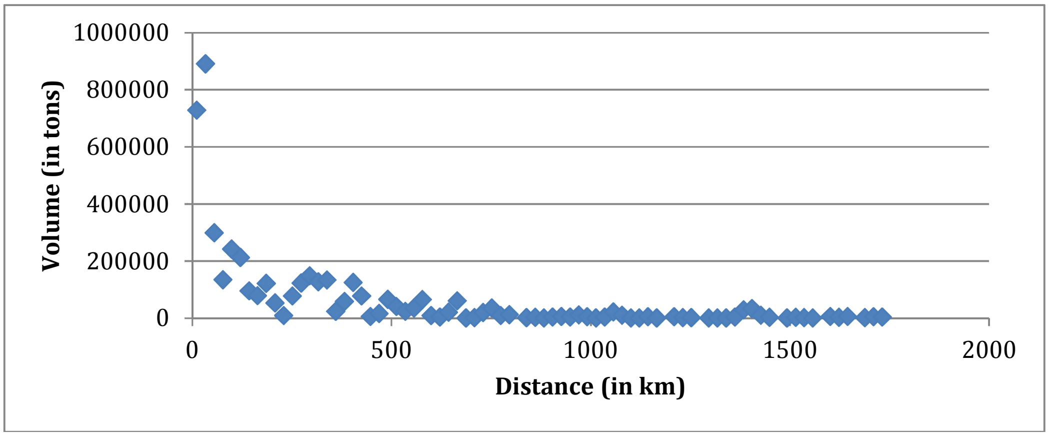


Figure 5.4: Observations for beverages binned in 80 bins

The established graphs, with the tons against the classified distance (as shown in Figure 5.4), form the final input for the derivation of decay functions.

### 5.6 Deriving decay functions from obtained data

In this section we will show how we generally derive decay functions from the obtained data, still using the commodity grouping beverages as an example.

The commodity grouping consists of alcoholic beverages, soft drinks and mineral water. In general these types of products have comparable logistical flows therefore the information of alcoholic beverages (obtained flow data) will be used as sample data for the derivation of the decay function. The data of the three biggest alcoholic beverage manufacturers in South Africa are combined (see Figure 5.2) and classified in 80 bins (see Figure 5.4). The next step is to fit a line (either a negative power or a negative exponential function) to the data that describes the decay in tons over distance in such a way that R-squared between the line and the data is minimized, resulting in the best fit. The (steepness of the) slope of the line is what is determined by the decay parameter, which in this case has a value of 1.677 (related to a negative power function, see Figure 5.5).

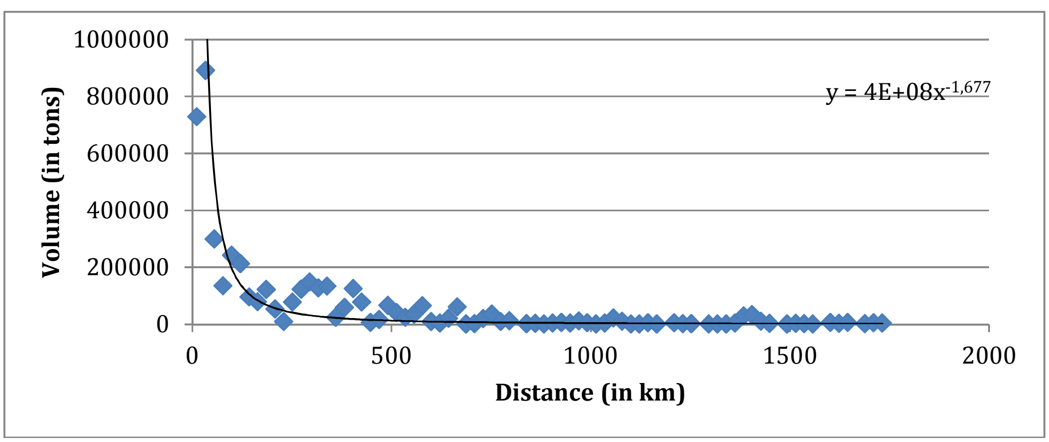


Figure 5.5: Fitting curve to obtained and binned data for beverages

The same procedure is followed for the other thirteen commodity groupings involved in the analysis. The resulting decay functions are listed in the table below. In Appendix I a table is presented with more detail regarding the binning of the data sets and the equivalent decay parameters related to the opposite decay function (negative power function into negative exponential function and visa versa).

| **Commodity grouping** | **Decay function** | **Decay parameter** |
|---|---|---|
| Beverages | Power | 1.68 |
| Bricks | Exponential | 0.01 |
| Cement | Exponential | 0.0068 |
| Fruit<br>• Deciduous fruit | <br>Power | <br>1.58 |
| • Citrus | Exponential | 0.0022 |
| • Sub-tropical | Power | 1.088 |
| • Viticulture | Exponential | 0.0034 |
| Jet fuel | Power | 1.63 |
| Motor vehicles | Power | 0.040 |
| Paper & paper products | Exponential | 0.0021 |
| Sugar Cane | Exponential | 0.052 |
| Wheat | Power | 0.58 |
| Cotton | Power | 0.75 |
| Tobacco | Power | 1.1 |

Table 5.2: Derived decay functions from obtained data

Now that we have decay functions that are derived from actual freight flow data, we are able to compare the new derived decay functions with the currently used decay functions to indicate the differences.

## 5.7 Newly derived vs. currently in use

In the Table 5.3 the newly derived decay functions are listed next to the decay functions currently in use. Some of the commodity groupings have shifted from a power function to an exponential function. This has to do with the fact that the decay function is chosen that fitted best (highest $R^2$) with the (binned) data. As mentioned earlier an overview of the equivalent decay function and parameters is given in Appendix I.

A first glance at both figures indicates big differences; the mean trip length (MTL) related to currently used decay functions is in almost all cases higher than the MTL related to the new derived decay functions. But also remarkable similarities; the figures for cement are nearly equal. As we have only one set of observations we do not have any insight in the spread of the parameters and therefore we cannot test if the decay functions differ significantly from each other.
When comparing the new derived decay parameters with the parameters used in comparable studies (see Appendix B) there seem to be more similarities. The agricultural products (the different fruits, ‘Tobacco’ and ‘Sugar cane’) are in line with the Dutch model (De Jong et al., 2010) as well as the models from Black (1973). The parameter value for ‘Motor vehicles’ is low, for ‘Paper and paper products’ it is average and for ‘Cement’ and ‘Beverages’ it is high. This is similar to Black’s models. The relative low value for ‘Bricks’ however contradicts the high value for ‘Structural clay products’ derived by Black. No reasonable explanation can be given at this stage.

| | New derived decay functions | | | Currently used decay functions | | |
|---|---|---|---|---|---|---|
| **Commodity grouping** | **Decay function** | **Decay parameter** | **MTL** | **Decay function** | **Decay parameter** | **MTL** |
| Beverages | Power | 1.68 | 286km | Power | 0.15 | 707km |
| Bricks | Exponential | 0.01 | 120km | Exponential | 0.035 | 65km |
| Cement | Exponential | 0.0068 | 310km | Exponential | 0.00644 | 318km |
| Fruit<br>• Deciduous fruit | Power | 1.58 | 723km | Power | 0.15 | 843km |
| • Citrus | Exponential | 0.0022 | 608km | Power | 0.15 | 928km |
| • Sub tropical | Power | 1.088 | 485km | Power | 0.15 | 662km |
| • Viticulture | Exponential | 0.0034 | 263km | Power | 0.15 | 329km |
| Jet fuel | Power | 1.63 | 202km | Power | 2.00 | 194km |
| Motor vehicles | Power | 0.040 | 607km | Power | 0.01 | 624km |
| Paper & paper products | Exponential | 0.0021 | 332km | Power | 0.15 | 615km |
| Sugar Cane | Exponential | 0.052 | 30,7km | Exponential | 0.10 | 3km |
| Wheat | Power | 0.58 | 553km | Power | 0.50 | 583km |
| Cotton | Power | 0.75 | 664km | Power | 0.15 | 668km |
| Tobacco | Power | 1.1 | 425km | Power | 0.15 | 1063km |

Table 5.3: New derived vs. currently used decay functions

However to be able to interpret the figures and give meaningful comments on both the differences and similarities it is also important to have an idea of the robustness of the decay functions.

*Robustness analysis*

This analysis should give us an idea how much the new derived decay functions deviate from the currently used decay functions from a practical point of view. In other words how many of the currently used decay parameters fall within an allowable range from the new derived decay parameter values. To be able to test the robustness of the derived decay functions we had to formulate an applicable definition. We define robustness as the change in the decay parameter given a certain change in the mean trip length (MTL). We use the MTL as it represents the transport distance as well as the volume transported.

As the MTL in many cases strongly depends on the geographical locations of the supply and demand points it is hard to predict the range of possible MTL values. For some commodity groupings (with a small range of possible flow sets and related MTLs) it might take several attempts to discover the allowable range of MTLs, which is very time consuming. Therefore we inversed the robustness analysis by changing the decay parameter values and looking at the effect on the MTL. Eventually this gives the same outcome for the analysis (see Figure 5.6).

The green and red line indicate respectively the upper and lower limit of the 10% deviation range (315 – 257km) of the obtained MTL (286km). The black dotted lines point out the related range for the decay parameter value (1.37 / 2.20). The chosen ranges are arbitrary and only used to give an indication. Therefore no (hard) conclusions can be based on the analysis.

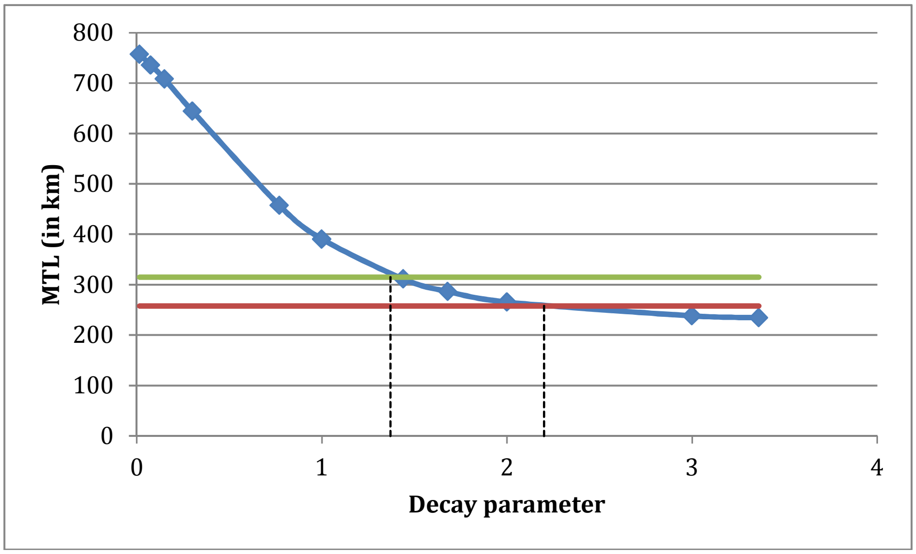


Figure 5.6: Robustness analysis of 'Beverages'

In the table below the range of possible decay parameter values is listed (fourth column) that lead to approximately the same MTL as the MTL related to the new derived decay parameters (range of +/- 10%, see Figure 5.6 as visualisation). In the last column the difference (in percentage) between the MTL related to the currently used decay parameter and MTL related to the new derived decay parameters can be found.

| **Commodity grouping** | **Decay function** | **Decay parameter** | **Range (+/- 10%)** | **Decay parameter in range** | **Deviation in MTL** |
|---|---|---|---|---|---|
| Beverages | Power | 1.68 | 1.37 / 2.20 | No | 147% |
| Bricks | Exponential | 0.01* | | | |
| Cement | Exponential | 0.0068 | 0.004 / 0.013 | Yes | 1% |
| Fruit<br>• Deciduous fruit | Power | 1.58 | 0.44 / > 3.16 | No | 16% |
| • Citrus | Exponential | 0.0022 | 0.001 / 0.007 | No | 53% |
| • Sub tropical | Power | 1.088 | 0.84 / 1.34 | No | 40% |
| • Viticulture | Exponential | 0.0034 | 0.001 / 0.014 | No | 65% |
| Jet fuel | Power | 1.63 | 1.2 / > 4.0 | Yes | 3% |
| Motor vehicles | Power | 0.040 | < 0.001 / 0.2 | Yes | 1% |
| Paper & paper products | Exponential | 0.0021 | 0.0018 / 0.0032 | No | 85% |
| Sugar Cane | Exponential | 0.052 | 0.04 / 0.06 | No | 22% |
| Wheat | Power | 0.58 | 0.43 / 0.75 | Yes | 6% |
| Cotton | Power | 0.75 | < 0.015 / > 3.0 | Yes | 1% |
| Tobacco | Power | 1.1 | 0.2 / > 3.3 | No | 11% |

Table 5.4: Robustness analysis for the new derived decay parameters

*Obtained data not suitable for this exercise

Black (1973) mentioned that in general the gravity model is not particularly sensitive to slight changes in its exponents. However from Table 5.4 we can see this does not count for all commodity groupings in our research as the robustness of the decay parameters differs significantly between the commodity groupings. In case of for example 'Tobacco' and 'Cotton' the decay parameters can be changed over a broad range without influencing the MTL significantly while for example the MTL values of 'Cement' and

‘Sugar Cane’ show considerable changes when the parameters are changed only slightly. In case of commodity groupings like ‘Tobacco’ and ‘Cotton’ with very limited supply and/or demand points (see Appendix G) there are not so many alternative flow sets possible. If there is for example only one supply point, all the demand points need to be supplied form this single supply point. In this case the value of the decay parameter does not influence the flows (as they are ‘fixed’) and therefore barely any change in the MTL can be observed.

The last two columns show that for certain commodity groupings like ‘Cement’ the currently used and new derived decay parameters are highly comparable while the parameters and related MTL of for example ‘Beverages’ differ significantly.

The findings from the robustness analysis and comparison between the current and new obtained decay parameters give rise to a new correlation analysis between decay and the possible influencing factors but this time based on just the decay parameters derived from the observed data.

## 5.8 Chapter summary

The proposed methodology to come to a regression model requires several reliable data sources. It is hard to verify all figures supplied by Conningarth but currently they offer the data that comes closest to data obtainable from a logistical survey or census and the presented figures appear appealing. As there was no data available to derive decay functions from we proposed to use the currently applied decay functions for the correlation and regression analyses. However a first univariate correlation analysis with these decay functions showed very weak and insignificant relations between the influencing factors and decay. Besides that, certain patterns in the decay functions are present, which makes them inappropriate for the purpose of regression analyses.

A comparison between the currently used decay parameters in South Africa and the decay parameters used in comparable models, revealed some interesting differences. These differences give rise to find out if the low correlation coefficients for the relations between the influencing factors and the decay from the initial univariate correlation analysis might have been caused by the incorrect assignment of decay functions to the commodity groupings. To test this hypothesis we have to repeat the correlation analysis with (for this purpose) newly derived decay functions. The differences between the two groups of decay parameters (current vs. new), underlined by the outcome of the robustness analysis, and the fact that the new derived decay parameters are more in line with the parameters used in comparable studies, strengthen the idea of performing a new correlation analysis. This correlation analysis only includes the 14 commodity groupings that have been assigned a new decay parameter. For the rest the new analysis will be similar to the initial one. In the next chapter the results from the new correlation analysis are presented.

# 6. Results

Now that we have decay functions based on actual freight flow data, we are able to repeat the initial correlation analysis to verify if the new decay parameters do show the expected relationship with the influencing factors. We will start this chapter with the outcome of the correlation analyses. The influencing factors that show a significant relationship with decay will be used in the regression analysis. This analysis will be presented in Section 6.2. We will conclude by discussing the verification and validation of the model in Section 6.3.

## 6.1 Correlation analysis

In Figures 6.1 and Table 6.1 the outcome of the correlation analysis between the newly derived decay parameters and the possible influencing factors (‘Value per ton’, ‘Scarcity’, ‘Supply concentration’ and ‘Local demand’) is presented. The analysis is based on the 14 commodity groupings presented in the previous chapter (and listed in Appendix G).

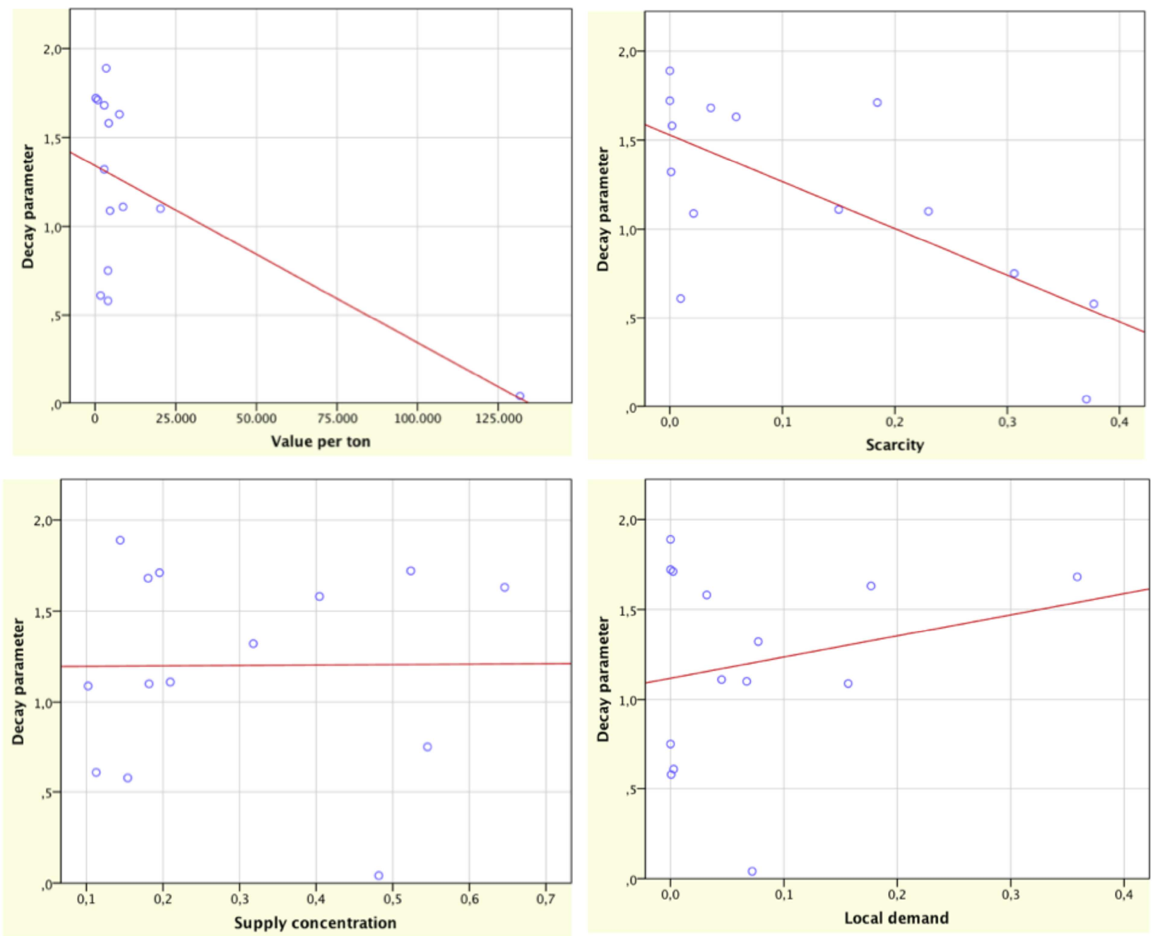



Figure 6.1: Correlation analysis between the new derived decay parameters and possible influencing factors*

*based on the 14 derived power functions listed in Appendix G

Again, like in Chapter 5, we had the choice between several approaches regarding decay functions. We could have split the 14 commodity groupings into two groups; one for commodity groupings for which the negative power function fitted best to the data (8) and one for commodity groupings for which the negative exponential function fitted best (6). However as14 commodity groupings is already a very small number of observations to base a regression model on, a further split up would make it even more difficult to draw founded conclusions. We could also have chosen to transform all negative power functions into negative exponential functions. This however would mean that we have to transform the decay functions for eight commodity groupings instead of six (when we would transform the other way around). Moreover we did the analysis for the currently used decay functions already based on a set of commodity groupings with negative power functions assigned. Therefore all 14 commodity groupings in this correlation analysis have been assigned a negative power decay function. This means that commodity groupings with a negative exponential function as a best fit with the obtained data have been assigned the closest related negative power functions (second best fit). In Appendix J the correlation coefficient tables of the correlations analyses based on the other approaches are presented. Some of the correlation coefficient tables show significant relations of moderate strength. However the squatter plots do not visually underline these relations.

The correlation coefficients related to the scatter plots in Figure 6.1 are summarized in the table below. The figures in the first row of the table are the correlation coefficients, the second row shows the observed significance levels (the chance that this relation is actually non-existing) and 'N' indicates the amount of observations included in the analysis.

| | | Value per ton | Scarcity | Supply concentration | Local demand |
|---|---|---|---|---|---|
| Decay parameter | Pearson Correlation | -,625 | -,692 | ,008 | ,217 |
| | Sig. (1-tailed) | ,008 | ,003 | ,489 | ,228 |
| | N | 14 | 14 | 14 | 14 |

Table 6.1: Correlation coefficient table (with new derived decay parameters)

*Verifying hypotheses*

Although we know that most of the new derived decay functions differ a lot from the currently used functions, the correlation analysis shows us that also these new decay functions do not clearly express the expected relationships with the influencing factors. Therefore we cannot yet conclude if (the values of the decay parameters in) the currently used decay functions are the reason for the weak relation between decay and the possible influencing factors as stated earlier in Chapter 5.

From the correlation coefficients in Table 6.1 we can see that 'Value per ton', 'Scarcity' and 'Local demand' have the correct sign given the hypotheses stated in Chapter 4. 'Supply concentration' seems to have no relationship with decay at all. Possible explanations for these observations will be considered during the discussion in the next chapter.

Given the observed significance levels (see Table 6.1) the weak to moderate relationships between 'Value per ton' and decay as well as 'Scarcity' and decay seem to be significant (using a significance level of 0.05 as stated earlier in Chapter 4).

When we look at the individual scatter plots (Figure 6.1) we see that especially the graph of 'Value per ton' and 'Supply concentration' gives rise to apply a logarithmic

transformation (see Section 4.4) to reduce the influence of the outlier (in both cases the commodity grouping 'Motor vehicles'). Therefore we perform a similar correlation analysis, as done before with the new derived decay functions however this time both the decay parameter as well as the scores of the influencing factors will be log transformed.
The coefficient table for this correlation analysis is presented below.

| | | Value per ton (log) | Scarcity (log) | Supply concentration (log) | Local demand (log) |
|---|---|---|---|---|---|
| Decay parameter (log) | Pearson Correlation | -,696 | -,389 | -,162 | ,082 |
| | Sig. (1-tailed) | ,003 | ,106 | ,290 | ,405 |
| | N | 14 | 12 | 14 | 11 |

Table 6.2: Correlation coefficient table (based on log transformed data)*

* The reduced amount of degrees of freedom are caused by missing or not transformable values

Although the log transformation increased the significance level of 'Scarcity' it decreased and therefore improved the significance level of both 'Value per ton' and 'Supply concentration'. The significance level of Supply concentration is still far above the threshold of 0.05. If we look at the scatter plot of the log transformed relation of Value per ton' and decay (below) we can see that the relation still does not appear to be very clear and strong and is still dominated by the outlier. This means that including (the log transformed) 'Value per ton' in a regression model to predict the decay parameter would be debatable and most probably meaningless (bad predictor). Therefore we decide to include only 'Scarcity', out of the four researched possible influencing factors above, in the regression model (see next section), as this factor showed a moderate relationship with decay.

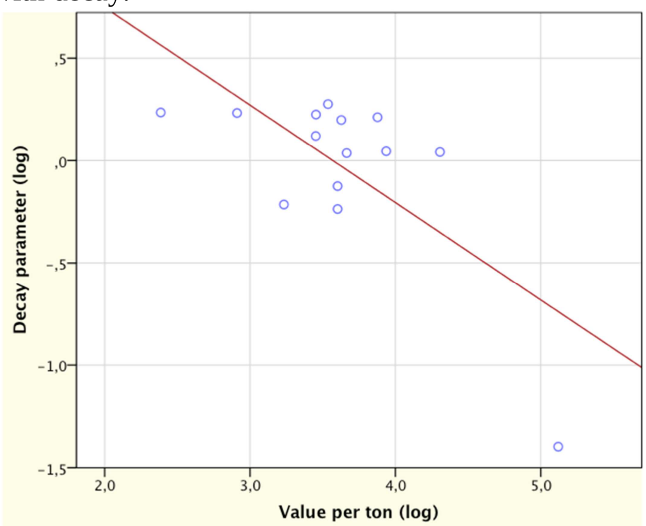



Figure 6.2: Scatter plot of the relation between the log transformed decay parameter and log transformed 'Value per ton'

Beside these possible influencing factors we have also studied two factors that are not scored based on known quantitative figures but based on the knowledge of expert; 'Homogeneity' and 'Brand'. We will now look into these two factors and their possible influence on the decay.

*Homogeneity & brand*
As stated in Section 4.3 we expect the level of homogeneity and the amount of brands combined with brand loyalty to influence the value of the decay parameters. The initial plan was to perform in-dept industry research on these two factors to be able to use figures derived straight from the industry itself (as hard figures are not readily available within CSCM). However due to the fact that the derivation of new decay functions got priority combined with time restrictions on the research forced us to use expert knowledge instead.
The scoring of the three factors (as ‘Brand’ exists of ‘Amount of brands’ and ‘Brand loyalty’) by the panel of experts is explained in detail in Appendix D.
These scores have been used in the correlation analyses as presented below.

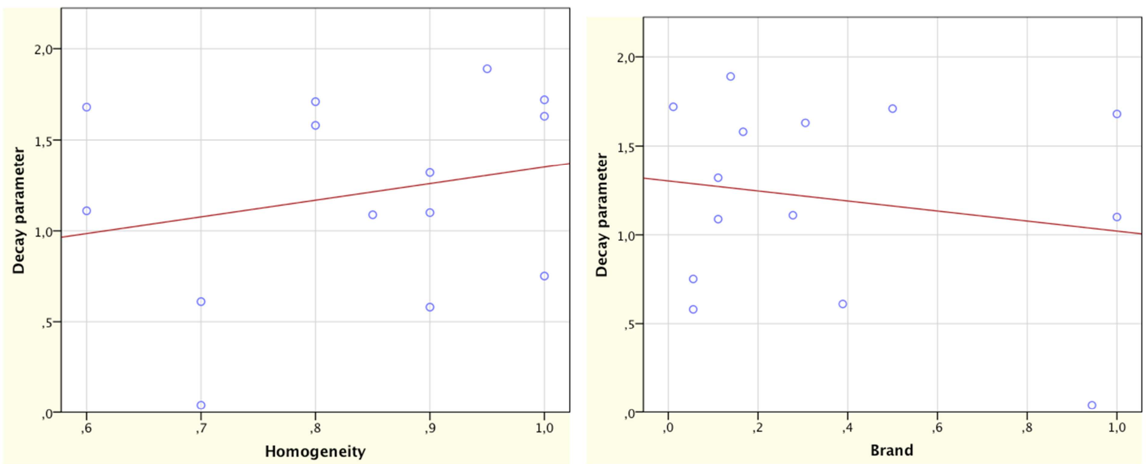



Figure 6.3: Correlation analyses between the new derived decay parameters, ‘Homogeneity’ and ‘Brand’

| | | Homogeneity | Brand |
|---|---|---|---|
| Decay parameter | Pearson Correlation | ,235 | -,186 |
| | Sig. (1-tailed) | ,210 | ,262 |
| | N | 14 | 14 |

Table 6.3: Correlation coefficient table

From the correlation coefficient table (as well as the scatter plots) we can see that both factors behave in line with the hypothesis stated in Section 4.3. The observed level of significance shows however that the relationship between the possible influencing factors and decay is weak. Both scatter plots do not give rise to use a log transformation as no clear outliers can be detected.
Given the analysis above there is no reason to include ‘Homogeneity’ or ‘Brand’ in a regression analysis. As ‘Scarcity’ turned out to be the only serious candidate for a regression analysis performing multivariate analysis is not necessary and there is no chance of multicollinearity. In the next section we will therefore present the outcome of the regression analysis with ‘Scarcity’ as only predictor.

### 6.2 Regression analysis

The main outcome of a regression analysis is normally given in three tables; a coefficient table, an ANOVA table and a table with the model summary. The coefficient table (Table 6.4) shows the coefficients for the regression model (listed under 'B') as well as the significance level (Sig.) for the included factors.

| Model | | Unstandardized Coefficients | | Standardized Coefficients | | |
|---|---|---|---|---|---|---|
| | | B | Std. Error | Beta | t | Sig. |
| 1 | (Constant) | 1,528 | ,148 | | 10,348 | ,000 |
| | Scarcity | -2,628 | ,791 | -,692 | -3,321 | ,006 |

Table 6.4: Regression coefficient table (decay parameter as dependent variable)

Given the coefficients in the table the regression model at hand is of the following form:

Decay parameter = 1.528 -2.628 Scarcity

Suggesting that the decay parameter for, for example, a 'Scarcity' of 0.4 would be 1.528 -2.628 * 0.40 = 0.46.

The significance level in the last column shows a different result as in the correlation analysis because SPSS performs standard a two sided test. This figure therefore needs to be divided by 2 and shows that the relationship between 'Scarcity' and decay is negative with a significance level of even less than 0.01.

From the analysis of variance (ANOVA) table we can find out how much of the total variability is being explained by the regression model (Regression Sum of Squares) and how much variability is still unexplained; the residual (or Residual Sum of Squares). In this case 1.864 out of 3.892 is being accounted for by the model. The other 2.028 is the amount of variability that still cannot be accounted for after the regression model has been fitted to the observations.

| Model | | Sum of Squares | df | Mean Square | F | Sig. |
|---|---|---|---|---|---|---|
| 1 | Regression | 1,864 | 1 | 1,864 | 11,029 | ,006[b] |
| | Residual | 2,028 | 12 | ,169 | | |
| | Total | 3,892 | 13 | | | |

Table 6.5: Analysis of variance (ANOVA) table (with the decay parameter as the dependent variable and a constant and 'Scarcity' as the predictors)

The model summary (Table 6.6) shows us the explanatory value of the regression model. R, the multiple correlation coefficient, is the correlation between the predicted and observed values, which is for a simple linear regression model equal to the correlation coefficient as found earlier in our correlation analysis. In the column of $R^2$ we can see the percentage of variability in the dependent variable that can be explained by the regression model. In this case the explanatory value of the model is 47.9%.

| Model | R | R Square | Adjusted R Square | Std. Error of the Estimate |
|---|---|---|---|---|
| 1 | ,692[a] | ,479 | ,435 | ,4110974 |

Table 6.6: Model summary (with the decay parameter as the dependent variable and a constant and 'Scarcity' as the predictors (indicated by [a]))

The explanatory value of the regression model is limited. Only half of the total variability in the decay parameter can be explained by applying this model. Therefore predicting the value of the decay parameter based on a regression model that only includes 'Scarcity' would probably result in significant deviations from reality and is for that reason not recommendable. However, the model can be used as an indication of the decay parameter value. A list of indication values for the decay parameter of the commodity groupings that have not been researched in-depth during this research can be found in Appendix K. However as accuracy of the regression model is low (see Section 6.3) we will not discuss all the listed decay parameters.

By looking at the scatter plot of 'Scarcity' vs. decay (Figure 6.1) the commodity grouping of bricks (at coordinates (0.009;0.61)) seems to be quite far of the trend line. We know from the description of the data (Appendix G) that they have been obtained in a different way (part of it was already binned into 6 bins) than most of the other commodity groupings. This might have influenced the derivation of the decay function and the value of the decay parameter and therefore the regression analysis.
Given the regression model and the assigned scarcity value for 'Bricks' this commodity grouping is supposed to have a decay parameter value close to 2. This is a high value relative to the decay parameters of other commodity groupings researched. However, if we compare this value with other studies, stone and clay products have often the highest scores (see Black (1972) for example). This indicates that our derived decay parameter value might be incorrect.
In case we would exclude 'Bricks' from the regression analysis the explanatory value of the model increases significantly ($R^2$ becomes 0.681). Besides the explanatory value also the significance level improved (0.001). This gives rise to a reconsideration of the obtained data for the commodity grouping of 'Bricks'.

## 6.3 Verification and validation

The next step after establishing the regression model is to verify and validate it.
A verified model means that the model operates correctly; it is logically consistent and complete.
Validation is the process of determining whether the model is a sufficiently accurate representation of the real system it is designed to reflect (Turnquist, 2006).

Without sufficient and correct data it is a difficult task to verify and validate the regression model. Because of its simplicity it is easy to say that it is logically consistent. However the moderate explanatory value of the model indicates that there should be more or other factors included and therefore we cannot state at this stage that the model is complete.
To be able to validate the model we need more decay functions derived from actual freight flow data (that reflects reality).

*Prediction intervals*

The 95%-prediction interval of the regression model gives an indication of the accuracy of the model itself. For every value of 'Scarcity' it gives the range of possible decay parameter values. This range is two times the standard error of the regression model (one to each side of the of a point on the regression line). If this range of values is small (small standard error) the accuracy of the model is high. However a broad interval (in combination with only a moderate relation) tends to allow almost all decay parameter values for every value of the explanatory variable. In Figure 6.4 the 95%-prediction interval for 'Paper & paper products' (with a scarcity value of 0.15) is shown.

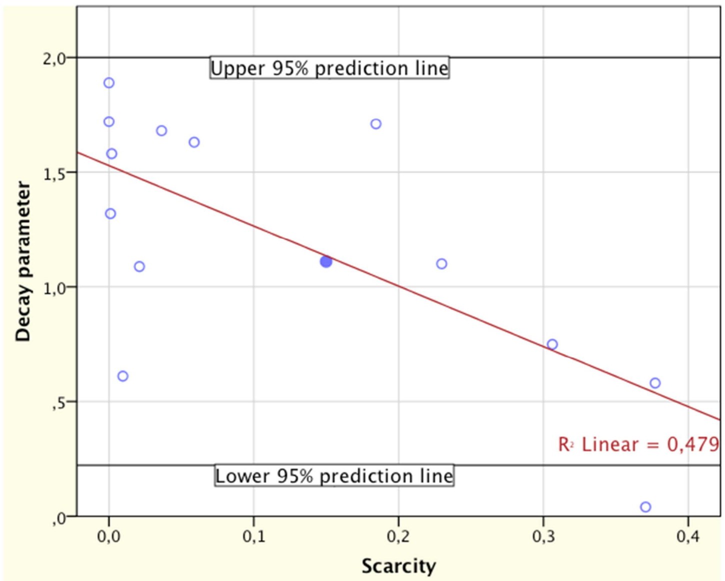



Figure 6.4: 95%-prediction interval for 'Paper & paper products

From the graph above it becomes clear that, as mentioned above, almost all possible decay parameter values (assigned to the commodity groupings in this research) are in the 95% range of the scarcity value of 0.15. This broad interval is caused by the relative big standard error of the regression model.

From a validation perspective this means that the accuracy of the model is rather low.

In the table below the 95%-prediction intervals for the other commodity groupings have been listed.

| **Commodity grouping** | **Lower 95% prediction line** | **Upper 95% prediction line** |
|---|---|---|
| Beverages | 0.49 | 2.37 |
| Bricks | 0.56 | 2.45 |
| Cement | 0.11 | 1.98 |
| Fruit<br>• Deciduous fruit | 0.57 | 2.47 |
| • Citrus | 0.57 | 2.48 |
| • Sub-tropical | 0.53 | 2.42 |
| • Viticulture | 0.58 | 2.48 |
| Jet fuel | 0.44 | 2.31 |
| Motor vehicles | -0.46 | 1.57 |
| Paper & paper products | 0.21 | 2.06 |
| Sugar Cane | 0.58 | 2.48 |
| Wheat | -0.49 | 1.56 |
| Cotton | -0.25 | 1.70 |
| Tobacco | -0.02 | 1.87 |

Table 6.7: 95%-prediction intervals for the regression model

*Road vs. rail*

The idea of using rail data to validate the model should be reconsidered at this stage. As mentioned before the currently used decay functions have been based partly on these rail data. From the comparison between the currently used and new derived decay functions we would expect the rail data to differ significantly from the data on road. Therefore a validation based on rail data would be inappropriate.

The rail data can however be used to obtain decay functions for rail (if it exist in suitable format) and perform a similar correlation analysis and maybe even a regression analysis. We could then compare the obtained decay functions, the relationships with the possible influencing factors and (if the correlation analysis gives rise to it) the regression model. This could give CSCM more insight in the difference in behaviour of the influencing factors between road and rail. This will be stated in the recommendations.

## 6.4 Chapter summary

A correlation analysis between the new derived decay functions and the possible influencing factors has been performed. The commodity groupings were for the purpose of the correlation analysis assigned a negative power decay function that fitted best to the related data. All relationships, except for the relation between decay and 'Supply concentration', follow the expected behaviour as stated in the earlier formulated hypotheses. However most relationships are weak and not significant except for the relationship between 'Scarcity' and decay. Therefore a regression analysis was performed with 'Scarcity' as the only predictor. The established regression model has an explanatory value of 47.9%, which can be increased to 68.1% if the commodity grouping 'Bricks', of which the data have been obtained in a different way compared to the other researched commodities, is left out of the analysis. The accuracy of the model is still not very high but it gives an indication of the value the decay parameter should have. In the next chapter possible explanations for the weak relationships between decay and the other possible influencing factors will be discussed.

# 7. Discussion

The fact that only one of the expected possible influencing factors seems to have a significant explanatory value for the decay parameters, gives rise to search for possible explanations for the non-existence of the other expected relationships. In this chapter different possible reasons will be presented and discussed.

In essence there are three main reasons for the problem at hand. Either the used data (sources) are incorrect (Section 7.1), the used methodology is inappropriate (Section 7.2) or the relations as demonstrated in literature do not fit the current characteristics of South Africa (Section 7.3). We will discuss the possible reasons in respective order below. We will end the chapter with an outlook into the future (Section 7.4).

## 7.1 Incorrect data (sources)

The whole study depends heavily on the data that are being used. Wrong input will obviously lead to wrong output. However as has been noted by many authors and experienced by CSCM too, high quality data are often hard to find and obtaining them is a costly and time-consuming exercise. If we look at the data that have been used to conduct the correlation and regression analyses we can distinguish three sources; Conningarth, expert knowledge and industry.

*Conningarth*

As mentioned in Chapter 5 the best readily available figures in South Africa regarding the characteristics of the commodity groupings at hand are at the moment produced by Conningarth. However this does not mean that the figures are correct. On almost weekly basis CSCM finds out about actual freight flows that do not match the data obtained from Conningarth. This has for example to do with bad alignment of what commodities are assigned to which grouping. Although this is a cycle of continuous improvement, mistakes in for example the locations of production or the amount or place of import will obviously have their influence on the assigned scores of the different researched possible influencing factors. The influence of these errors might for certain commodity groupings be significant.

*Expert knowledge*

Expert knowledge is a great source of information that is rather easy to obtain but it has its down sides. The data are only as good and extensive as the knowledge of the experts and often has some level of subjectivism. In an ideal case the expert knowledge would be used in combination with a fact-based research or broad survey. Both were unfortunately not possible in this study due to time and budget restrictions. The used scores for the factors 'Homogeneity' and 'Brand' therefore need to be used with a certain level of caution. Although experts claim to apply inter-subjectivism (representing the subjectivity of a broad group) it is hard to say what the actual quality of these figures is, as mentioned fact-based industry research should be done to verify this data source.

*Industry*

Given the fact that the currently used decay functions have been found by trial-and-error methods and sophisticated estimations ('guesstimations') the quality of the decay functions obtained by in-depth industry research may expect to be higher. However also this in-depth industry research has not been perfect. In many cases the industry figures have been based on only one or a few companies. This might have led to skewed data in case a company was specialized in short or long haul transport, a specific region or a specific sub-commodity. The percentage of obtained data might for some commodity groupings have been too low to obtain a reliable and representative sample resulting in comparable errors as just mentioned. In some cases the obtained data needed extra unit – weight or value – weight conversions. As we need volume versus distance figures for the derivation of decay functions (see Figure 5.2) an amount of cars or the value of cars needs to be converted into volume (tons). These are often estimations or averages as specific conversion rates are not known (even by the providers of the data). Another important factor was the assignment of high-level supply and demand location data to specific low-level MD's (or cities). Because of confidentiality some of the companies provided consumption data only on a sub-provincial level, while MD levels were needed for distance calculations. All these conversions and adaptations will have influenced the accuracy of the data. The significance of this influence is hard to quantify.

## 7.2 Inappropriate methodology

Next to incorrect data also an inappropriate methodology can be the cause of the non-existence of the correlation between decay and the possible influencing factors. Hereby one should think of how the metrics have been chosen, the way the commodities have been grouped, the application of model specific distance choices and we can even question if the right possible influencing factors have been chosen.

*Choice of metrics*

The way the metrics have been chosen has a major influence on the scoring of the possible influencing factors. The metrics for 'Value per ton', 'Supply concentration' and 'Local demand' have been chosen based on literature and are therefore expected to show similar correlations but the metrics for 'Scarcity', 'Brand' and 'Homogeneity' have been designed specifically for this research. It is very well possible that a different metric, with different scoring would have led to a stronger, more significant correlation.

*Grouping of commodities*

As has been mentioned in Chapter 1 and 5 the way commodities are aggregated into commodity groupings might influence the values of the decay parameters as well as the scoring of the commodity groupings on the different possible influencing factors. Black (1972) studied the differences between decay parameters applied in two different models. One model used 24 commodity groupings while the second model applied 80 (see Section 2.4). The outcome of his study was that there was a clear difference in the magnitude of the decay parameters (the spread in the decay parameter values of the more aggregated commodity groupings was far smaller). However the decay parameter values of the two models were proportional to each other. For example in both models stone-related commodity groupings were assigned the highest decay parameter values (11.25 and 5.325) and industrial machinery the lowest (-.05 and 0.25). This suggests that the correlation would not be effected by the grouping of the commodities.

However we have no information about differences in the scoring of the influencing factors. The fact that different commodities with different characteristics (for example different purposes, qualities or values per ton) have been aggregated into the same commodity groupings might have led to compound and blurred scores. Especially the scores for 'Value per ton' and 'Supply concentration' will be influenced. The scores of the first one for obvious reasons and the scores of the latter one because supply points might be specialised in the production/supply of a certain product (type). For example in case of many supply points one would expect short transport distances however if the supply points all produce a different kind of product (type) this expected behaviour will not be observed. But as mentioned in the first chapter model complexity will increase significantly if more commodity groupings would be added. Compared to other models the distinction between 64 commodity groupings is already of higher detail than the average model and does not deviate much from the 80 commodity groupings studied by Black. Therefore only a change in the way the commodities are grouped (the composition) would actually be an option for investigation.

*Selection of commodity groupings*

The selection of the commodity groupings included in this study could be of significant influence in the applied correlation analyses. In case of the value of the goods our selection of groupings is very skewed to the low value goods with 'Motor vehicles' as the exception. This makes it difficult to verify the hypothesis regarding the relationship between decay and 'Value per ton'. Because no data were made available by the industry, none of the mining commodity groupings have been part of the analysis, while they have been included in comparable studies (like the one carried out by Black (1972)). These commodity groupings generally have a high supply concentration (limited amount of mines) and high local demand (processed close to the mines). Their value per ton has a broad range (coal vs. diamonds). Therefore the exclusion of mining might have a significant influence on the scores of the possible influencing factors.

*Size of studied regions*

Although the level of detail in which the study area is divided might not really influence the value of the decay parameter it will definitely influence the scores of the influencing factors. The South African model uses 356 areas (MD's). Black (1972) studied only nine regions on a national level in the US (not clear how the size of a local production area is defined), which are therefore expect to be larger than the MD regions. In case of local demand (total consumption of a certain product in one region divided by total production of that product in the same region) the size of the area that is considered as local might have been much bigger in the US study than this study in South Africa. Therefore one would expect far lower scores for local demand in South Africa (currently ranges only between 0 and 0.4). Dividing South Africa in fewer regions will keep the decay parameters more or less the same but it will definitely influence (increase) the scores for 'Local demand'.

*Data processing*

Another important point in this discussion is the way obtained data have been processed. In the previous section we already mentioned the conversion of for example units into tons but as the obtained data often included different steps in the transportation process, finding the right distance for deriving the decay functions was complicated and needed processing of the data. The data included for example transport distances from the production plant to the package facility, from the package facility to the warehouse and from the warehouse to the final customer (all provided as separated shipments). This is in contrast with the model applied by CSCM that only cares about the distance from the initial production to the final customer. As explained for 'Motor vehicles' in Appendix G a lot of steps were necessary to trace back the goods from the final customer to their initial production point. This might have influenced the derived decay parameter values.

*Distance metric*

We can also question the use of the distance metric as mentioned in Section 2.2. In this study actual road distance has been used while an alternative like Euclidian distance, used by Blake (1972), might have led to different values (more negative decay parameters as distance is reduced). To find out if the influence of the metric is a cause for the difference in correlation strength and significance between this study and literature, an extra study would be necessary.

However as the model eventually needs to translate the generated freight flows onto an actual road network (with actual transport distances), it would be impractical to use a different distance metric for the derivation of decay functions.

The use of a different metric should therefore result in significant model improvements (this should be the outcome of the additional study) before implementation will be considered.

*Choice of possible influencing factors*

Before we even had to choose the metrics for the influencing factors we first had to decide on the possible influencing factors that we wanted to include in this study. Partly based on literature and partly based on well-founded hypothesis. This however does not mean that we have been complete. There might be other factors that correlate with decay stronger and more significant than the ones proposed in this study.

*Choice of decay function*

In the correlation analyses we have chosen to apply negative power decay functions. We could, as explained in earlier chapters, also have chosen for a different approach. We could have assigned all commodity groupings a negative power decay function or split the set of commodity groupings in two groups; one for which a negative power decay function fitted best to the related data and one for which a negative exponential function fitted best to the data. However for the currently used decay functions we have seen that neither the assigning of negative exponential functions nor the split up did result in better correlations (see Appendix E). In case of the new derived decay functions the assigning of negative exponential functions to all commodity groupings did not result in better correlations (see Appendix J). The split up between the two groups on the other hand seem to show stronger correlations between decay and the possible influencing factors. However the scatter plots are not really convincing and the amount of observations per group is really small. Therefore the split up would only be an interesting option if the amount of observations (commodity groupings with sufficient actual freight flows) would increase.

Moreover from Appendix I we can see that the R-squared values for the best fitting decay function and their alternative do not differ much (same order of magnitude). This suggests that it should not make a real difference what approach is applied.

*Binning*

In Appendix H we have already discussed the influence of binning data, which seems hard to quantify. In the next chapter we propose an alternative research plan that does not include a binning procedure at all.

### 7.3 Non-fit with current South African characteristics

The studies regarding decay parameters done in the past took place in developed countries like the US, England and the Netherlands. It seems a reasonable question if the findings for the other countries apply to South Africa given the fact that geographical, economical and cultural differences are significant.

*Decay parameters*

Black (1972) showed that resizing the studied area did not have significant influence on the derived decay parameters. Therefore the difference in size between the countries being researched is not expected to seriously influence the correlation. On the other hand the fact that the majority of freight transportation in the country takes place just between Johannesburg and the two other biggest cities (Cape Town and Durban) might have significant influence on the decay functions as is shown in the following example.

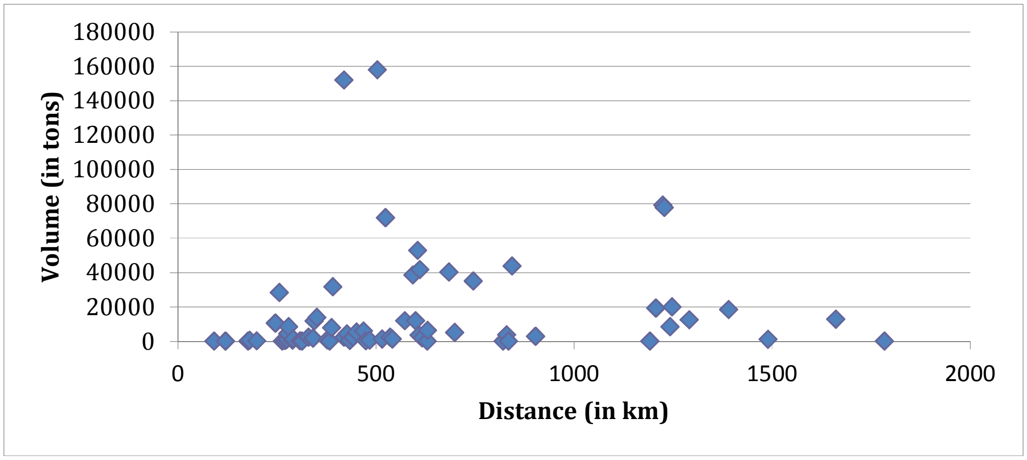



Figure 7.1: Observations of ‘Paper & paper products’

From the graph above it is clear that there are two peaks at 500km and 1200km. These are exactly the distances between Durban and Johannesburg and Johannesburg and Cape Town (freight flows between Cape Town and Durban are going via Johannesburg). This indicates that the main paper mill is located close to Johannesburg.

This effect is reduced by binning the data but might reduce even more (or in a different way) if we would use a more aggregated area level. Currently the MD level is used. If we would use a more aggregated province level instead, all transport also gets aggregated. Peaks in the graphs will be less visible and influencing, obviously at the cost of loss of amount of observations.

*Time factor*

Obviously a lot has changed globally over the past forty years. As suggested by for example Helvig (1964) the exponent varies with time reflecting improvements in transport, changes in production and inventory locations, changes in demand for transportation, varying mode preferences and changes in area interaction. This is an important reason to make the comparison between South Africa now and the US and England four decades ago with caution. As no recent studies have been carried out that confirm that relationships found in the past are still valid in the current economical and infrastructural environment, it is hard to make firm statements without further research.

*Value per ton*
Because of differences between the countries and time periods, the performed studies might have dealt with very different costs of transport. This could be an influencing factor in the relationship between 'Value per ton' and decay. Fuel prices were lower as well as labour and vehicle costs. On the other hand product prices were also lower, which makes it hard to indicate the influence of relative costs without a proper research on this subject.
Another issue that might explain the difference in travel distance between high and low value goods has to do with a trend on the trade market. Since the last decade it has become popular to use agricultural goods as speculation instruments. This has influenced the storage and (therefore) the transported distance significantly. Especially the lower value agricultural products have increased in their transported distance, resulting in less negative decay parameters.

## 7.4 Outlook into the future

The current trend where the business environment is changing into a more supply chain oriented environment can have significant influence on the current flows (Turnquist, 2006). The traditional 'within' firm decisions are changing into 'across' firm decisions resulting in new flows that transcend the usual and more or less predictable flows of single separate commodity groupings. Raw materials (for example 'Wood & wood products') may for example travel longer distances as the related processed product (in this case 'Paper & paper products') production facilities are located closer to the end consumer causing lower total transport costs. This means that decay parameters in decay functions of raw materials will decrease while the decay parameters in decay function of processed or end product (often associated with a higher price) will increase. Contradicting for example the hypothesis of the relationship between 'Value per ton' and decay.

Another evolving dimension in trade flows, globalisation with e-commerce as example, will also have its effect on the freight transportation flows. Suddenly the amount of (information about) supply points increases tremendously (as the market becomes global), because of this international competing countries start specializing and in many cases this means that more products will be imported. This will increase the scarcity of many but mainly processed or end products (from Asia, the US and Europe). The logistical chains get shorter as often products are stored and shipped immediately from the production plant or a big central warehouse. Transport is often outsourced to specialised parties who are, because of economy of scale, able to ship the products of different clients more efficient. This will reduce the total average transportation distance of mainly the end products and therefore influence the (historical) relationship between 'Value per ton' and decay.

For the current form of the model the impact of e-commerce will be limited as most original supply and demand point stay the same (except for the increasing importance of seaports). However as soon as CSCM decides to implement more logistical elements (warehouse locations, shipper behaviour, etc.) the increasing influence of e-commerce on the transportation flows should be accounted for.

## 7.5 Chapter Summary

Many factors have or may have influenced the correlation and regression analyses. Especially the level of detail (size of the regions), selection of commodity groupings, amount and quality of the obtained data and the processing of the data have had their impact on the values of the decay parameters and scores of the possible influencing factors. Besides the issues with data and methodology also the different time frames and countries will have influenced the comparison between the studies, the derived decay functions and their relationships with the possible influencing factors.

The different possible reasons for the non-existence of correlation between decay and possible influencing factors suggest improvements in the current research design as well as completely new studies. In the next chapter we will besides the overall conclusions also discuss these recommendations.

# 8. Conclusions and recommendations

This research started based on a main research question. Now that all the sub questions have been answered in the previous chapters the main question can be answered. This will be followed by general conclusions based on the performed analyses and related discussion (Section 8.1). Finally recommendations regarding model development and further research will be presented (Section 8.2).

## 8.1 Conclusions

### *8.1.1 The main research question*

The question that gave rise to this research was formulated and presented in the first chapter of this report as follows:

*How can the current decay functions of the commodity freight flow model applied on the freight flows of South Africa transported by road be improved in a way that both the current mapping of the freight flows as well as the forecasts of the freight flows will be improved by focussing on better insights in the factors influencing decay?*

The answer to this question can be given by combining the findings from the previous chapters and is stated below.

By applying a regression model based on the relationship between new derived decay functions and 'Scarcity' a prediction with moderate accuracy can be made for the values of the decay parameters for all commodity groupings in South Africa transported by road.
The use of these predicted decay functions should improve current mapping as well as forecasting the freight flows. However therefore validation of the model by newly obtained data is necessary.

The fact that the ratio between import and total demand ('Scarcity') can be used to predict (or influense) the decay function means that (radical) changes in this ratio will have his reflexion on the decay function. If it can be foreseen that the ratio will change in the (near) future the decay parameter in the forecasting module should be adjusted accordingly. This will improve the accuracy of the forecasts.

### *8.1.2 Data*

*Current decay functions*

The correlation analysis between the currently used decay functions and the possible influencing factors (derived from literature) indicated no significant relationships. Based on the related scatter plots one could conclude that similar decay functions were assigned to commodity groupings with very different scores on the possible influencing factors.

A comparison between the current and new derived decay functions shows big differences. However as only one set of observations is available (all used for the derivation of the new decay functions) nothing can be said about the spread in variation of the values of the new derived decay functions. Therefore no conclusions can be made regarding the significance of the differences between the two sets of decay functions.

*Robustness of the new decay functions*
If one would accept a +/- 10% deviation from the obtained mean trip length (MTL), only five out of fourteen currently used decay functions lay within the allowable range. This gives enough reason to apply a correlation analysis with the new derived decay functions.

### *8.1.3 Correlation & regression*

The observed relations between the new derived decay parameters and the six possible influencing factors were all in line with the hypothesis as stated on beforehand. However only ‘Scarcity’ has a relationship with decay that is significant and of moderate strength.
The regression model with only ‘Scarcity’ included as predictor has the following form:

Decay parameter = 1.528 -2.628 Scarcity

The explanatory value $R^2$ is 0.479 at a significance level of 0.05. Even at a significance level of 0.003 the relation between ‘Scarcity’ and decay is expected to be negative. If the commodity grouping ‘Bricks’ is excluded from the regression analysis (based on the obtained data and comparison with decay functions for stone or clay products in other models) the $R^2$ increases to 0.681 with a significance level of less than 0.001.

As no verification data are available it is hard to quantify the quality of the model. Given the 95% prediction intervals the accuracy of the model is not very high. Therefore it can in the best case be used as an indication for the decay parameter value.

### *8.1.4 Limitations*

The research has many limitations of which quantity and quality of available data have been the most important factors. Further has the used methodology played an important role as for example the size and amount of the regions under research have had serious influence on the scores of some of the possible influencing factors. The same counts for the choice of the commodity groupings included in our research. All these choices have influenced the scores and therefore the correlation and regression analyses.
However it is impossible to conclude that the limitations mentioned are the cause of the weak and insignificant relationships between decay and the possible influencing factors as several of the expected relationships were based on studies performed over more than forty years ago and in other countries. As both decay functions and the scores of the possible influencing factors vary over time and between countries the comparison should be made with caution.

Nevertheless this research has revealed some serious issues in the application of non-fact-based gravity modelling, indicated points that can and need to be taken into account in future derivation of decay functions and pointed out some interesting starting points for further research. These will be presented in the next section.

## 8.2 Recommendations

There are many recommendations to give regarding the application of the South African commodity freight flow model, derivation of decay functions in the future and starting points for further research. In this respective order we will briefly present the most important recommendations below.

*Model improvements*

Currently the value of the model is limited to trade flows while there are many ambitions and attempts within CSCM to translate these flows into transportation flows of trucks without appropriate model adaptations. The first recommendation is to try to improve the model by integrating a transportation flow module (including transportation and inventory cost components) into the current trade flow based freight flow forecasting model. This will improve the link with actual freight flows on the road network. In line with this model extension it would be recommendable to include (shipper/driver based) behavioural conditions in the road choice. This would next to the fact that it makes the model more realistic make it also easier to use truck counts as a proper quality check tool, something that is completely missing at the moment.

In the current forecasting module all parameters have fixed values. By implementing an accurate regression model at least changes in the parameter value can be predicted based on assumed changes in the explanatory variable (in this case 'Scarcity').

Although it is obvious that CSCM is not able to check all figures that come from third parties (like Conningarth) it would be advisable to build in some quick quality checks on the data. The model output is only as good as the data that are used as input.

*Future derivation of decay functions*

In the quest to find reliable decay functions, obtaining sufficient and high quality data is the absolute key factor. Without new data it is impossible to validate and improve the regression model, a model that could fulfil the wish of scientifically (or fact-based) derived decay functions.
As long as no obligation exists for companies to document details about their transportation, the best way to obtain the necessary data would be by carrying out a survey or census. However as mentioned these methods cost a lot of time and money so convincing government or other financial strong partners to support such a project should be a priority. Without these sources data need to be obtained straight from the source and that has turned out to be a very difficult and especially a time-consuming task.

However it should not be to hard to keep the companies that have supplied data for this research involved in the coming years by sharing results from the analyses. For them it is often interesting to see how far their products actually move, how that changes over time and what the prediction for the future is.
With obtained data over several years a comparison can be made between the derived decay functions for each year to see if significant changes occur over time. This type of data (five succeeding years) is already available for the commodity grouping 'Cement', which could therefore be used as a pilot.

If more data are obtained, it is possible to aggregate commodity groupings so that model and decay function behaviour can be compared with other models and their parameters (that often have higher levels of commodity aggregation). This should lead to more insights in the behaviour of both and therefore in time higher model accuracy.

As mentioned in Chapter 1 does Flowmap only allow power and exponential decay functions as model input. It might be interesting to study the influence of applying a compound function in the derivation of decay functions and the correlation and regression analyses. If this results in a regression model with a higher explanatory value it might be worth to discuss the implementation of this functionality in Flowmap.

*Future research*
The origin and destination specific data obtained for the derivation of the decay functions can also be used to test the quality of the commodity freight flow model. To be able to use the data for quality check purposes it is necessary to aggregate the output flows to the same level (province or sub-province for example) as the provided data.
In this way it is also possible to compare the differences in outcome between the currently used and new derived decay functions.

With the available data relationships between decay and other possible influencing factors could be tested. Instead of changing the factor it is also possible to just test alternative metrics.

The obtained data from the commodity grouping ‘Bricks’ should be reconsidered. Probably an alternative way to obtain the data should be designed and implemented.

The reasoning behind the hypotheses claiming relationships between ‘Homogeneity’ and decay and ‘Brand’ and decay seems still plausible. Therefore we suggest that a fact-based industry research should be carried out to verify and improve the figures obtained from expert knowledge. Moreover extending the panel of experts (with more industry authorities for example) would also increase the quality of the figures.

*Alternative research setup*
With the knowledge gathered throughout this research it became possible to design a new research proposal that will take away many difficulties experienced on the way.

As mentioned before the most important step is to obtain as much OD-data for every commodity grouping as possible. Then for every commodity grouping construct OD-tables based on the obtained data. Next produce per commodity grouping a set of OD-tables based on a gravity model using different decay parameters with the restriction that the totals of the rows and columns of these tables need to be equal to (or within a certain range from) the row and column totals of the OD-tables based on the obtained data. Compare the individual cells of the observed OD-table and the constructed OD-tables and calculate the sum of the squared differences ($R^2$). Select the gravity model with the decay parameter that results in the lowest $R^2$. This gravity model can now be used to produce OD-tables based on the supply and demand tables supplied by Conningarth.

This alternative method should be able to give better estimations for the decay functions as they are derived straight from the obtained data and no binning procedures are included. However the major challenge in this method will be to obtain sufficient OD-specific data. To simplify the exercise the amount of regions should be seriously reduced. Constructing OD-tables on a provincial level (9x9) should be sufficient. If the gravity model is not able to come to an OD-table because the observed table has too many empty cells (not enough data are available) one might aggregate commodity groupings that are closely related.

With the data obtained during the current research it would not have been possible to carry out the proposed alternative method as in many cases the data included only distance and volume and no origins or destinations were included, which are essential to construct the OD-tables in this alternative method.

*New applications*
In the future this type of research in combination with the proposed model extensions (inclusion of transportation flows) can be a great starting point of new research. As mentioned in Chapter 2 interest in emissions and other environmental issues is increasing. Based on reliable data (or a validated regression model) it should be possible to predict the influence of possible new behaviour changing policies from the government on $CO_2$ emissions from road transport for example. This and many more interesting studies one could think of can evolve from the research concentrating on decay functions.

# References

Anderson, J.E., (1979): A Theoretical Foundation for the Gravity Equation. American Economic Review, 69(1), March, 106-116

Black, W.R. (1971): The utility of the gravity model and estimates of its parameters in commodity flow studies. Proceedings of the association of American geographers. Vol. 3, 28-32

Black, W.R. (1972): Interregional commodity flows: some experiments with the gravity model. Journal of regional science. Vol. 12. 107-118

Black, W.R. (1973): An analysis of gravity model distance exponents. Transportation, Vol. 2(3), 299-312

Bröcker, J., Rohweder H.C. (1990): Barriers to international trade, Methods of measurement and empirical evidence. Annals of Regional Science, Vol 24(4), 289-305

Bröcker, J., Korzhenevych, A., Riekhof, M. (2010): Predicting freight flows in a globalising world. Research in Transportation Economics, Vol. 31(1), 37-44

Cascetta, E., Di Gangi, M. (1996): A multi-regional input-output model with elastic trade coefficients for the simulation of freight transport demand in Italy. In proceedings of the 24th PTRC Summer Annual Meeting, England

Celik, H.M., Guldmann, J.M. (2007): Spatial interaction modeling of interregional commodity flows. Socio-Economic Planning Sciences, Vol. 41(2), 147-164

Chen, T.M., et al. (2005): Report on model specification and calibration results TRANS-TOOLS (TOOLS for TRansport forecasting ANd Scenario testing) Deliverable3. Funded by 6th Framework RTD Programme, TNO Inro, Delft

De Jong, G., Gunn, H., Walker, W. (2004): National and international freight transport models: an overview and ideas for future development. Transport Reviews, Vol. 24(1), 103-124

De Jong, G., Ben-Akiva, M. (2007): A micro-simulation model of shipment size and transport chain choice. Transportation Research Part B, Vol.41, 950 - 965

De Jong, G., De Bok, M., Ruijs, K., Wentink, D. (2010): Schatting BasGoed. Significance

Du Potgieter, L.J., Nanny, R., Van Zyl, S.J.J. (1993): A Guide to the Fifth Edition of the Standard Industrial Classification of all Economic Activities (SIC). Research Report No. 245

Ferrari, M.J., Bjørnstad, O.N., Partain, J.L., Antonovics, J. (2006): A Gravity Model for the Spread of a Pollinator-Borne Plant Pathogen. The American Naturalist, Vol. 168(3), 294-303

Flowerdew, R., Aitkin, M. (1982): A method of fitting the gravity model based on the Poisson distribution. Journal of Regional Science, Vol. 22, 191-202

Fotheringham, A.S., O' Kelly, M. (1989): Spatial interaction models: formulations and applications. Kluwer Academic Publishers (Boston)

Frankel, J.A., Wei, S.J. (1998): Regionalization of world trade and currencies: economics and politics. The University of Chicago Press

Gujarati, D.N. (2006): Essentials of econometrics, third edition. McGraw – Hill

Hamilton, C.C. (1983): Die Suid-Afrikaanse goederevervoermark. Special report for the National institute for transport and road research, CSIR, South Africa

Hanson, M.E. (1966): Project METRAN: An Integrated Transportation System for Urban Areas. M.I.T. Press

Haynes, K.E., Fotheringham, A.S (1984): Gravity and spatial interaction models. Scientific Geography Series, Vol. 2

Heien, P.M. (1968): A note on log-linear regression. Journal of the American Statistical Association, Vol.63, 1034-1038

Helvig, M. (1964): Chicago's external truck movements – spatial interaction between the Chicago area and its hinterland. University of Chicago Press

Hensher, D.A., Button, K.J. (2008): Handbook of transport modelling, ($2^{nd}$ ed.). Elsevier, 715-727

Iacono, M., Krizek, K.J., El-Geneidy, A. (2008): Access to destinations: how close is close enough? Estimating accurate distance decay functions for multiple modes and different purposes. Access to destinations study for Minnesota Department of Transportation

Jennrich R.I., Sampson P.F. (1976): Newton-Raphsoandn Related Algorithms for Maximum Likelihood Variance Component Estimation. Technometrics, Vol. 18(1), 11-17

Linneman, H. (1966): An econometric study of international trade flows. North-Holland Publishing Company, Amsterdam

McNally, M.G. (2000): The four-step model. Handbook of transport modelling, ($2^{nd}$ ed.), Chapter 3 (35-52). Elsevier

Ortuzar, J. d. D., Willumsen, L. G. (2001): Modelling transport (3rd ed.). John Wiley

Östlund, B., Janson, G., Ruijgrok, C., Tavasszy, L., Petersen, M. S. (2003): A Logistics Module for SAMGODS. Swedish Institute for Transport and Communications Analysis, Stockholm

Pretorius, A.M.J. (1990): Inter-reginal transport and energy modelling in South Africa: A brief review of past research and future directions. National institute for transport and road research, CSIR, South Africa

Research unit for transport economic and physical distribution studies (1990): Freight transport data bank report, 2nd edition

Roy, J.R., Thill, J. (2004), Spatial interaction modelling. Regional Science, Vol 83(1-2), 339-361

Smith, W.J.J. (1971): A quantitative study of road haulers and ancillary road transport users in the republic of South Africa 1971-72.

Southworth, F. (2006): Defining Future Needs. Freight Demand Modeling; Tools for Public-Sector Decision Making (Conference of the Transportation Research Board)

Taaffe, E.J., Howard, H.L. Gauthier, O'Kelly, M.E. (1996): Geography of Transportation, (2nd ed.). Prentice Hall

Tavasszy, L.A. (1994); Characteristrics and capabilities of Dutch freight transportation systems models; RAND report MR-382-EAC/VW, RAND Europe, Leiden.

Tavasszy, L.A., Van de Vlist, M., Ruijgrok, C., Van de Rest, J. (1998): Scenario-Wise Analysis of Transport and Logistic Systems with a SMILE. Presented at 8th World Conference on Transport Research, Antwerp, Belgium, 1998.

Tavasszy, L.A. (2006): Freight Modelling – An overview of international experiences. TRB Conference on Freight Demand Modelling: Tools for Public Sector Decision Making, Washington DC, 2006

Timmermans, H.J.P. (1996): A stated choice model of sequential mode and destination choice behaviour for shopping trips. Environment & Planning A, Vol.28, 173-184

Tobler, W., Wineberg, S. (1971): A Cappadocian Speculation. Nature, Vol. 231, 39-41

Turnquist, M.A. (2006): Characteristics of effective freight models. Freight Demand Modeling; Tools for Public-Sector Decision Making (Conference of the Transportation Research Board)

Van der Zwan, J., Van der Wel, R., De Jong, T.,Floor, H. (2005): Manual Flowmap 7.2. Faculty of Geographical Sciences, Utrecht University, The Netherlands

Williams, H.C.W.L. (1976): Travel demand models, duality relations and user benefit analysis. Journal of Regional Science, Vol. 16(2), 147-166

Williams, I. (2002): Feasibility Study of SCGE Models of Goods Flows in Sweden. SIKA & TNO-Inro

Yang, C., Regan, A.C., Tae Son, Y. (2009): Another View of Freight Forecasting Modeling Trends. KSCE Journal of Civil Engineering (2010), Vol. 14, 234-242

# Appendices

## Appendix A List of commodity groupings and currently used decay functions

| # | Commodity grouping | Decay function | Decay parameter |
|---|---|---|---|
| 1 | BARLEY | Power | 0.50 |
| 2 | COTTON | Power | 0.15 |
| 3 | DECIDUOUS FRUIT | Power | 0.15 |
| 4 | CITRUS | Power | 0.15 |
| 5 | SUBTROPICAL FRUIT | Power | 0.15 |
| 6 | VITICULTURE | Power | 0.15 |
| 7 | GRAIN SORGHUM | Power | 0.50 |
| 8 | LIVESTOCK (SLAUGHTERED) | Power | 0.01 |
| 9 | MAIZE | Power | 4.00 |
| 10 | SOYA BEANS | Power | 0.50 |
| 11 | SUNFLOWER SEED | Power | 0.50 |
| 12 | VEGETABLES | Power | 0.01 |
| 13 | WHEAT | Power | 0.50 |
| 14 | POULTRY PRODUCTS | Power | 0.01 |
| 15 | DAIRY | Power | 0.01 |
| 16 | SUGAR CANE | Exponential | 0.10 |
| 17 | OTHER AGRICULTURE | Power | 0.50 |
| 18 | COAL MINING | Exponential | 0.05 |
| 19 | CRUDE PETROLEUM & NATURAL GAS | Exponential | 0.01 |
| 20 | IRON ORE (HEMATITE) | Power | 2.78 |
| 21 | MAGNETITE | Power | 0.50 |
| 22 | CHROME | Power | 2.04 |
| 23 | COPPER | Power | 0.50 |
| 24 | MANGANESE | Exponential | 0.00 |
| 25 | TITANIUM | Power | 0.50 |
| 26 | ZINC | Power | 0.50 |
| 27 | OTHER NON-FERROUS METAL MINING | Power | 0.50 |
| 28 | STONE QUARRYING, CLAY & SAND-PITS: GRANITE | Exponential | 0.04 |
| 29 | STONE QUARRYING, CLAY & SAND-PITS: LIMESTONE & LIME WORKS | Exponential | 0.04 |
| 30 | STONE QUARRYING, CLAY & SAND-PITS: OTHER | Exponential | 0.04 |
| 31 | MINING OF CHEMICAL & FERTILIZER MINERALS | Exponential | 0.04 |
| 32 | OTHER NON-METALLIC MINERALS | Power | 0.50 |
| 33 | OTHER MINING | Power | 0.50 |
| 34 | FOOD AND FOOD PROCESSING | Power | 0.15 |
| 35 | BEVERAGES | Power | 0.15 |
| 36 | TOBACCO PRODUCTS | Power | 0.15 |
| 37 | TEXTILES, CLOTHING, LEATHER PRODUCTS AND FOOTWEAR | Power | 0.15 |
| 38 | WOOD AND WOOD PRODUCTS | Power | 0.50 |
| 39 | FURNITURE | Power | 0.15 |
| 40 | PAPER & PAPER PRODUCTS | Power | 0.15 |
| 41 | PRINTING AND PUBLISHING | Power | 0.01 |
| 42 | INDUSTRIAL CHEMICALS | Power | 0.15 |
| 43 | FERTILIZERS AND PESTICIDES | Power | 2.00 |
| 44 | PHARMACEUTICAL, DETERGENTS AND TOILETRIES | Power | 0.15 |
| 45 | PETROLEUM REFINERIES AND PRODUCTS OF PETROLEUM/COAL | Power | 2.00 |
| 46 | RUBBER PRODUCTS | Power | 0.15 |
| 47 | OTHER CHEMICALS | Power | 0.15 |
| 48 | NON-METALLIC MINERAL PRODUCTS | Power | 0.20 |
| 49 | BRICKS | Exponential | 0.04 |
| 50 | CEMENT | Exponential | 0.01 |
| 51 | FERROCHROME | Power | 0.01 |
| 52 | FERROMANGANESE | Power | 0.01 |
| 53 | OTHER IRON AND STEEL BASIC INDUSTRIES | Exponential | 0.04 |
| 54 | NON-FERROUS METAL BASIC INDUSTRIES | Power | 2.00 |
| 55 | METAL PRODUCTS EXCLUDING MACHINERY | Power | 0.08 |
| 56 | MACHINERY AND EQUIPMENT | Power | 0.15 |
| 57 | ELECTRICAL MACHINERY | Power | 0.15 |
| 58 | MOTOR VEHICLES | Power | 0.01 |
| 59 | MOTOR VEHICLE PARTS AND ACCESSORIES | Power | 0.01 |
| 60 | TRANSPORT EQUIPMENT | Power | 0.20 |
| 61 | OTHER MANUFACTURING INDUSTRIES | Power | 0.15 |
| 62 | WATER SUPPLY | Power | 0.15 |
| 63 | GAS | Power | 2.00 |
| 64 | JET FUEL | Power | 2.00 |

## Appendix B Decay parameter values from comparable models

| # | Commodity grouping | Decay parameter | Correlation |
|---|---|---|---|
| 1 | Agriculture | 1.602 | 0.680 |
| 2 | Processed food | 0.912 | 0.771 |
| 3 | Petroleum | 2.752 | 0.978 |
| 4 | Crude oil | 1.363 | 0.945 |
| 5 | Ores | 0.887 | 0.984 |
| 6 | Metals & semi conductors | 0.539 | 0.637 |
| 7 | Crude minerals | 1.462 | 0.855 |
| 8 | Fertilizers | 0.969 | 0.670 |
| 9 | Chemicals | 2.148 | 0.817 |
| 10 | Others | 1.016 | 0.815 |

Table B.1 Decay parameter values obtained and used in BasGoed model (Dutch)
Data source: De Jong, G., De Bok, M., Ruijs, K., Wentink, D. (2010): Schatting BasGoed. Significance

| Shipper Group | Title | Decay parameter | Correlation |
|---|---|---|---|
| 1 | Meat and Dairy Products | 2.625 | 0.957 |
| 2 | Canned and Frozen Foods | 2.825 | 0.960 |
| 3 | Candy, Beverages and Tobacco Products | 1.975 | 0.963 |
| 4 | Textile Mill and Leather Products | 0.850 | 0.995 |
| 5 | Apparel and Related Products | 1.025 | 0.961 |
| 6 | Paper and Allied Products | 1.500 | 0.977 |
| 7 | Basic Chemicals, Plastics and Synthetics | 1.675 | 0.970 |
| 8 | Drugs, Paints and Other Chemical Products | 2.450 | 0.991 |
| 9 | Petroleum and Coal Products | 0.275 | 0.856 |
| 10 | Rubber and Plastics Products | 0.950 | 0.981 |
| 11 | Lumber and Wood Products | 0.675 | 0.946 |
| 12 | Furniture and Fixtures | 0.975 | 0.952 |
| 13 | Stone, Clay and Glass Products | 5.325 | 0.980 |
| 14 | Primary Iron and Steel Products | 0.950 | 0.983 |
| 15 | Primary Nonferrous Metal Products | 1.100 | 0.961 |
| 16 | Fabricated Metal Products | 1.875 | 0.959 |
| 17 | Metal Cans, Misc. Fab. Metal Products | 1.525 | 0.981 |
| 18 | Nonelectrical Industrial Machinery | 0.250 | 0.985 |
| 19 | Machinery (except Electrical and Industrial) | 0.600 | 0.970 |
| 20 | Communication Products and Equipment | 0.425 | 0.982 |
| 21 | Electrical Products and Supplies | 0.375 | 0.990 |
| 22 | Motor Vehicles and Equipment | 0.500 | 0.992 |
| 23 | Transportation Equipment | 0.400 | 0.966 |
| 24 | Instruments, Photograpic Equipment, Watches and Clocks | 0.500 | 0.928 |

Table B.2 Classification based on 24 major Shipper groups (Black, 1972)
Data source: US Census of Transportation, 1967, “Commodity Transportation Survey”, Reports TC67-Cl-1 through TC67-Cl-24, issued 1970

| TCC Group | Title | Decay parameter | Correlation |
|---|---|---|---|
| 201 | Meat, Poultry and Small Game Products | 1.55 | 0.926 |
| 202 | Dairy Products | 5.25 | 0.995 |
| 203 | Can and Preserved Fruits, Vegatables and Seafoods | 0.70 | 0.963 |
| 204 | Grain Mill Products | 2.25 | 0.937 |
| 206 | Sugar Beet and Cane | 1.75 | 0.937 |
| 207 | Confectionary and Related Products | 0.30 | 0.962 |
| 208 | Beverages and Flavroing Extracts | 2.60 | 0.975 |
| 209 | Miscellaneous Food Preparations and Products | 2.50 | 0.965 |
| 221 | Cotton Broadwoven Fabrics | 0.50 | 0.992 |
| 222 | Man Made Fiber and Silk Broadwoven Fabrics | 0.35 | 0.968 |
| 227 | Carpets, Rugs, Mats, Textile | 0.60 | 0.960 |
| 228 | Yarn and Thread | 0.50 | 0.998 |

Table B.3 Classification based on 80 distinguished commodity groupings (Black, 1972)
Data source: US Census of Transportation, 1967, “Commodity Transportation Survey”, Part II

| TCC Group | Title | Decay parameter | Correlation |
|---|---|---|---|
| 229 | Miscellaneous Basic Textile | 0.50 | 0.928 |
| 231 | Men's, Youths' and Boys' Clothing | 0.95 | 0.906 |
| 233 | Women's, Misses', Girls' and Infants' Clothing | 0.30 | 0.929 |
| 239 | Miscellaneous Fabricated Textile Products | 1.05 | 0.903 |
| 242 | Lumber, Dimension Stock and Other Mill Products | 0.75 | 0.963 |
| 243 | Millwork and Prefabricated Wood Products | 0.35 | 0.962 |
| 249 | Miscellaneous Wood Products | 1.25 | 0.879 |
| 251 | Household and Office Furniture | 1.30 | 0.959 |
| 262 | Paper | 1.15 | 0.940 |
| 263 | Paperboard, Fiberboard and Pulpboard | 1.20 | 0.973 |
| 264 | Converted Paper and Paperboard Products | 1.35 | 0.950 |
| 265 | Containers, Boxes and Related Products | 3.25 | 0.986 |
| 281 | Industrial Chemicals | 1.65 | 0.973 |
| 282 | Plastic Materials and Plasticizers | 1.00 | 0.899 |
| 283 | Drugs | 0.25 | 0.915 |
| 284 | Doap and Detergents, Cleaning Preparations | 1.45 | 0.950 |
| 285 | Paints, Varnishes, Lacquers | 1.85 | 0.956 |
| 287 | Agricultural Chemicals | 3.40 | 0.993 |
| 289 | Miscellaneous Chemical Products | 1.50 | 0.971 |
| 291 | Products of Petroleum Refining | 0.20 | 0.857 |
| 295 | Paving and Roofing Materials | 3.30 | 0.989 |
| 301 | Tires and Inner Tubes | 0.95 | 0.958 |
| 306 | Miscellaneous Fabricated Rubber Products | 1.00 | 0.979 |
| 307 | Miscellaneous Plastic Products | 0.90 | 0.951 |
| 314 | Footwear (except Rubber) | 0.45 | 0.900 |
| 316 | Luggage and Handbags | 0.25 | 0.897 |
| 322 | Glass and Glassware, Pressed and Blown | 2.05 | 0.974 |
| 324 | Hydraulic Cement | 11.25 | 0.949 |
| 325 | Structural Clay Products | 3.20 | 0.976 |
| 326 | Pottery and Related Products | 0.60 | 0.892 |
| 327 | Concrete, Gypsum and Plaster Products | 3.30 | 0.992 |
| 329 | Miscellaneous Nonmetallic Mineral Products | 2.00 | 0.981 |
| 331 | Steel Works and Rolling Mill Products | 0.90 | 0.980 |
| 332 | Iron and Steel Castings | 2.15 | 0.979 |
| 333 | Nonferrous Metals, Primary Smelter Products | 1.05 | 0.921 |
| 335 | Nonferrous Metal, Basic Shapes | 0.55 | 0.960 |
| 336 | Nonferrous and Nonferrous Base Alloy Castings | 1.20 | 0.990 |
| 339 | Miscellaneous Primary Metal Products | 1.35 | 0.987 |
| 341 | Metal Cans | 2.85 | 0.982 |
| 342 | Cutlery, Hand Tools and General Hardware | 0.50 | 0.944 |
| 343 | Plumbing Fixtures and Heating Appartus | 0.65 | 0.956 |
| 344 | Structural and Miscellaneous Metal Products | 2.30 | 0.949 |
| 345 | Bolts, Nuts and Other Industrial Fasteners | 1.60 | 0.986 |
| 346 | Metal Stampings | 0.70 | 0.994 |
| 348 | Miscellaneous Fabricated Wire Products | 1.45 | 0.939 |
| 349 | Miscellaneous Fabricated Metal Products | 1.00 | 0.950 |
| 351 | Engines and Turbines | 0.40 | 0.983 |
| 352 | Farm Machinery and Equipment | 0.65 | 0.984 |
| 353 | Construction, Mining Machinery | 0.35 | 0.975 |
| 354 | Metalworking Machinery and Equipment | 0.45 | 0.992 |
| 355 | Special Industry Machinery | 0.75 | 0.945 |
| 356 | General Industrial Machinery and Equipment | -0.05 | 0.970 |
| 357 | Office, Computing and Accounting Machines | 0.00 | 0.975 |
| 358 | Service Industry Machines | 0.90 | 0.910 |
| 359 | Miscellaneous Machinery and Parts | 1.25 | 0.993 |
| 361 | Electrical Transmission and Distribution Equipment | 0.10 | 0.935 |
| 362 | Electrical Industrial Apparatus | 0.45 | 0.987 |
| 363 | Household Appliances | -0.05 | 0.967 |
| 364 | Electric Lighting and Wiring Equipment | 0.50 | 0.928 |
| 365 | Radio and Television Receiving Sets | 0.45 | 0.986 |
| 366 | Communication Equipment | 0.05 | 0.947 |
| 367 | Electronic Components or Accessories | 0.35 | 0.961 |
| 369 | Miscellaneous Electrical Machinery and Equipment | 1.35 | 0.972 |
| 371 | Motor Vehicles and Motor Vehicle Equipment | 0.45 | 0.988 |
| 372 | Aircraft and Parts | 1.90 | 0.901 |
| 379 | Miscellaneous Transportation Equipment | 1.40 | 0.907 |
| 382 | Measuring and Controlling Instruments | 0.70 | 0.918 |
| 386 | Photographic Equipment and Supplies | 0.05 | 0.985 |

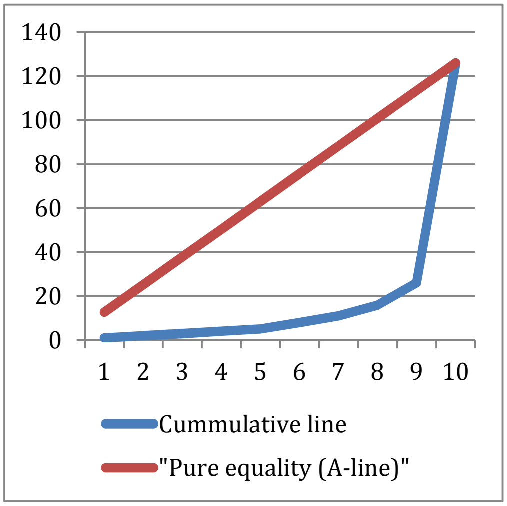
Table B.3 (Continued) Classification based on 80 distinguished commodity groupings (Black, 1972)

## Appendix C Background for using Gini coefficients

The Gini coefficient, founded by the Italian statistician Corrado Gini in 1912, is a measure of the inequality of a distribution, a value of 1 expressing total equality and a value of 0 maximal inequality (see below). The expression for the Gini coefficient (G) is as follows:

G=A/(A+B)

A is the surface below the cumulative line and A and B together represent the total surface below the “Pure equality”-line.



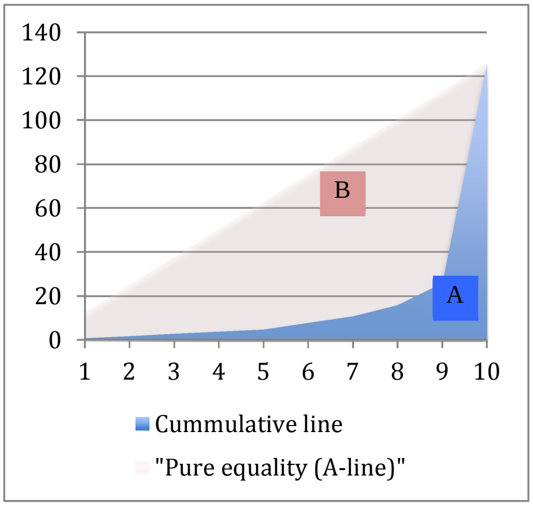




Figure C.1A and Figure C.1B Visual representation of the Gini coefficient

The graphs above show the visual representation of a Gini coefficient related to brands in a commodity grouping. In this example 50% of the sub-commodities has only 1 brand, 20% has 3 brands, 10% has 5 brands, 10% has 10 brands and the last 10% has 100 brands. This results in a Gini coefficient of 0.78, associated with a quite strong level of equality.

The Gini coefficient can as explained in Section 4.3 be used to get better insight in the influence of the inequality within a commodity grouping (different characteristics of sub-commodities) on the scores of the possible influencing factors.
If for example there are a few sub-commodities with little tonnage transported but an exorbitant high product value relative to the other sub-commodities the Gini coefficient will be close to 1. This might indicate why the commodity grouping has a very high average value but does in logistical sense not behave as expected. In this case the commodity grouping might only travel short distances for example.

An in-depth industry research is necessary to obtain the essential data to be able to calculate (or approximate) the Gini coefficient.

## Appendix D Scoring of 'Homogeneity' and 'Brand'

In the table below the scores assigned to 'Homogeneity' and 'Brand' are listed.

| Commodity groupings | Homogeneity | Brand | # of Brands | Brand loyalty | Brand score |
|---|---|---|---|---|---|
| Beverages | 0.6 | 1.0 | 100 | 80 | 180 |
| Bricks | 0.7 | 0.39 | 30 | 40 | 70 |
| Cement | 0.8 | 0.05 | 40 | 50 | 90 |
| Decidious fruit | 0.8 | 0.17 | 20 | 10 | 30 |
| Citrus fruit | 0.9 | 0.11 | 10 | 10 | 20 |
| Sub tropical fruit | 0.85 | 0.11 | 10 | 10 | 20 |
| Viticulture | 0.95 | 0.14 | 10 | 15 | 25 |
| Jet fuel | 1.0 | 0.31 | 5 | 50 | 55 |
| Motor vehicles | 0.7 | 0.94 | 80 | 90 | 170 |
| Paper & paper products | 0.6 | 0.28 | 30 | 20 | 50 |
| Sugar Cane | 1.0 | 0.01 | 1 | 1 | 2 |
| Wheat | 0.9 | 0.06 | 5 | 5 | 10 |
| Cotton | 1.0 | 0.06 | 5 | 5 | 10 |
| Tobacco products | 0.9 | 1.0 | 80 | 100 | 180 |

Table D.1 Scoring of 'Homogeneity' and 'Brand'

In scoring both factors we have taken into account that it should be possible to include other commodity groupings when sufficient data are obtained. In case of 'Homogeneity' we have chosen to base the lowest score on the commodity group of processed food, which is not yet included in our research, and score the rest of the commodity groupings accordingly. Therefore the lowest score on 'Homogeneity' that is included in our research is for beverages and paper & paper products.

To score 'Brand' we have first scored 'Number of brands' and 'Brand loyalty' on a scale of 0 – 100. Next we have added up the two scores and normalized these scores between 0 and 1.

**Appendix E Correlation analyses based on currently used decay functions**

In Chapter 5 the correlation analysis was based on the 53 commodity groupings that initially had been assigned a negative power decay function. In the tables below the results of correlation analyses with the other 11 commodity groupings (that have been assigned a negative exponential function) and all 64 commodity groupings together (the commodity groupings with a negative exponential function have been assigned the best fitting negative power function) are presented.

| | | Decay parameter | Value per ton | Scarcity | Supply concentration | Local demand |
|---|---|---|---|---|---|---|
| Decay parameter | Pearson Correlation | 1 | -,003 | -,401 | -,141 | -,156 |
| | Sig. (1-tailed) | | ,496 | ,111 | ,340 | ,323 |
| | N | 11 | 11 | 11 | 11 | 11 |

Table E.1 Correlation coefficient table of the correlation between decay and possible influencing factors *

* Based on the 11 commodity groupings that were initially assigned a negative exponential decay function

| | | Decay parameter | Value per ton | Scarcity | Supply concentration | Local demand |
|---|---|---|---|---|---|---|
| Decay parameter | Pearson Correlation | 1 | -,184 | -,181 | ,148 | -,167 |
| | Sig. (1-tailed) | | ,073 | ,076 | ,121 | ,093 |
| | N | 64 | 64 | 64 | 64 | 64 |

Table E.2 Correlation coefficient table of the correlation between decay and possible influencing factors **

** Based on all 64 commodity groupings, assigned a negative power decay function

None of the correlation analyses show a strong or significant relationship.

## Appendix F Effect of resizing regions on 'Supply concentration'

As mentioned in Chapter 5 the size of the regions used for this study are different from the sizes of the regions used by for example Black (1972). The correlations tested are partly based on Black's earlier research. The size of the regions may therefore by of influence and a cause for the non-existence of the correlation between the currently used decay functions and the possible influencing factors.
In the table below the outcome of the correlation analyses based on the 11 provinces, instead of the 356 MD's, in South Africa are presented.

| | | Supply concentration | Supply concentration (resized) |
|---|---|---|---|
| Decay parameter | Pearson Correlation | ,182 | ,180 |
| | Sig. (1-tailed) | ,096 | ,099 |
| | N | 53 | 53 |

Table F.1 Correlation between currently used decay parameters and 'Supply concentration' based on province level (11 regions)*

* Based on 53 commodity groupings initially assigned a negative power decay function

| | | Supply concentration | Supply concentration (resized) |
|---|---|---|---|
| Decay parameter | Pearson Correlation | ,008 | ,234 |
| | Sig. (1-tailed) | ,489 | ,210 |
| | N | 14 | 14 |

Table F.2 Correlation between new derived decay parameters and 'Supply concentration' based on province level (11 regions)**

** Based on 14 commodity groupings with new derived decay parameters (all assigned a negative power decay function)

The sign of the correlation coefficients in both tables indicate a (very weak) relationship that is in contrast with the hypothesis in Section 4.3.

## Appendix G Overview of selected commodity groupings

In this appendix an overview of the commodity groupings is given that have been selected for the derivation of new decay functions. Insight is given in the composition of the commodity grouping, their industry, the obtained data and the processing of the data.

**Motor vehicles**

The commodity grouping of motor vehicles (1.1 million tons) exist for the far majority out of cars. About 40% of all new motor vehicles bought in South Africa are being imported by the seaports of Durban, Port Elisabeth and East London. The other 60% is being produced by the 8 car manufacturers in South Africa; Toyota (22%), Volkswagen (21%), GMSA (15%), Ford (8%), Nissan (7%), Mercedes (7%), BMW (6%) and Renault (3%).

The cars that get exported (278.000 tons) are mostly produced close to the seaports where they leave the country. The seaports used for exports are the same as the ones used for importing cars.

Some of the car manufacturers transport their cars to the dealers themselves but most of the cars are transported by logistical companies (Grindrod (16%), Motorvia (42%), VDS(42%)).

Difficulties in data translation

Before an overview is given regarding the obtained data it is necessary to express one of the difficulties that needs to be coped with when using these data. In many cases the logistical companies only keep track of the shipments they fulfil and not of the whereabouts of specific cars. The obtained information contains therefore every step that a car makes in the logistical chain between production/import facility and final customer as a separate shipment. In other words the trips from plant to intermediate depot and intermediate depot to final dealer are seen as separate shipments by the logistical companies. If we would use this information without processing on beforehand the average travel distance would be much shorter than in reality. Therefore all the cars that arrive on an intermediate depot need to be reallocated to the different original plants or ports based on the proportion of total inflows from that plant/port.

Although the cars are being transported over longer distances (port-city – Johannesburg + Johannesburg-final destination) the freight flow model is, as mentioned in Chapter 3, based on initial origin and final destination only. Therefore we reassign the vehicles to their initial origin. If we would have taken the actual transport distance it would have increased the average travel distance and therefore forced the model to flow cars between origins and destinations with a longer distance. It would for example force cars demanded in Zimbabwe to be transported from Port Elisabeth (1200km) instead of from Durban (1000km). Moreover the general trend is that more and more the direct, shortest route is being chosen between origins and destinations instead of making use of a far-off-route depot or warehouse. Therefore using the shortest route is consedered more appropriate.

Information obtained:

VDS provided a year round overview of their car transportation activities. Over 90% of their transportation is between the port-cities and Johannesburg. Of those destinations Johannesburg is the only hub (intermediate depot). So the cars that come from Johannesburg originate from one of the port-cities. These cars are therefore reallocated

to the different port-cities based on the proportion of inflows into Johannesburg (as mentioned above).
VDS uses an average weight of 1.85 tons per vehicle.

Motorvia provided a six months overview of their car transportation activities. They confirmed the assumption that multiplying the data by two would give a good estimation of year round figures. The origins are specifically indicated; Durban and Weswood (Gauteng). All the cars dispatched from Weswood do originally come from Durban. Therefore we have added all destinations and related volumes to the figures of the Durban dispatches. The destinations are indicated on a sub-province level (provinces are divided in 1 to 5 sub regions). Middle points per sub-province have been taken as final destination. The average weight used internally at Motorvia is 1.6 tons per vehicle.

Grindrod provided a four months overview of their car transportation activities. The volumes are in units but internally they use an average weight of 1.5 ton per vehicle. Both origins and destinations are given on a level of provinces however as the production facilities of the car manufacturers are known the origins can be traced back to specific locations. For the destinations middle points of the provinces have been taken.
From the cars that Grindrod transports from Durban to Johannesburg the far majority (90%) is meant for the Gauteng market. Only 10% get dispatched further. The four months figures have been multiplied by three to make them year round figures and comparable with the figures of the other companies.

**Beverages**
The commodity grouping of beverages (14.4 million tons) exist out of alcoholic beverages as well as soft drinks and mineral water.
Only half a million tons is imported through the seaports of Durban, Port Elisabeth and Cape Town. The rest is produced all over the country.
The market of alcoholic beverages is divided between three big players (SAB Miller (58%), Distell (30%) and Brandhouse (10%)) and a couple of small ones. They all take care of their own logistics. The soft drink market as well as the market for mineral water is dominated by Coca Cola. Their transport is done by ABI.

Information obtained:
Distell provided a detailed overview of their logistical operations for a year round period (based on 2010 figures). This includes their bulk transport, internal transport, wines and packed liquor. Their bulk and internal transport is limited to very short distances to production and bottle facilities. Therefore only the packed liquor figures, from the production and bottle facilities to the final depots, are taken into account. The places of origin and destination are given on a town/city level so only distance between the OD-pairs needed to be calculated.

SAB Miller provided three matrices (total amount of loads, total amount of kilometres and total tonnage) for their primary (brewery – warehouse) and logistical flows. The 7 breweries are placed in the columns and the 59 depots in the rows. By dividing the total amount of kilometres driven between each OD-pair (brewery –depot combination) by the total amount of loads delivered from the brewery to the depot the average trip length between the brewery and the depots can be determined. This average distance is assigned to the volume from the brewery to the depot.

Brandhouse provided year round figures in a comparable format as the packed liquor of Distell with the only difference that the short internal flows were still present. These distances have been filtered out to overcome double counting. As well as for Distell this means incurring a small distance error.

From all three beverage suppliers only primary transport (brewery to final depot) is provided. The secondary transport (depot to final client) takes place in a 50 to 100-kilometre radius. Including the secondary transport would have increased the average distance and the spread. The transport volumes provided include for all three suppliers their packaging material (about 15% of the total volumes).

The information of ABI has been promised but never received. But as the logistical structure of the soft drinks/mineral waters and alcoholic beverages is comparable the received information as described above should give a sufficient sample for the commodity grouping.

**Bricks**

The commodity grouping of bricks (15.9 million tons) consist of face as well as pavement bricks, tiles (all made out of clay) and many other mostly marble related products. The products within this commodity grouping get produced and consumed all over the country. Bricks represent the majority of the transported tonnage within this group. Bricks are mainly produced by a couple of big players (Corobrick, Westend, African bricks, Ocon) and a lot of smaller players.
The tile industry is less than a tenth of the brick industry. They are mainly produced by Ceramics (310.000 tons) and Johnson (200.000 tons). Italtile owns the majority (65%) of the selling points (under the name CTM) all over the country. Part of their tiles is produced locally by the two SA producers the other part is being imported.

Obtained information:
In corporation with the Clay Brick Association we did a small survey within her group of members (producing 80% of all bricks in South Africa). The members were asked to give the percentage of their bricks that gets transported over a certain distance (six categories). In the end we did receive response from all the members. Therefore the survey is definitely representative for the total transport of bricks in South Africa.
Ceramics provided year round data from their tile transportation. The origins and destinations are divined on a municipality level. Only distance needed to be calculated.
Johnson provided high-level figures regarding the radius of distance of their transport. These figures could be used to verify the figures from Ceramics.

Although the survey covers the far majority of the brick industry it has some serious limitations. The data are already binned in only six bins therefore we lose a lot detail and it is difficult to compare the output data in an accurate way. A second disadvantage is that it is hard to include the obtained data related to the tile industry as their figures are not grouped at all. Comparable bin sizes can be used but it will never be really comparable.

**Cement**

The commodity grouping of cement (16.2 million tons) exist of ready mix for cement and concrete and cement and concrete based products like building blocks, building sections, bricks and tiles.

The main producers of cement and concrete in South Africa are PPC (40%), Afrisam (30%) and Lafarge (20%).

About 3 million tons of cement and concrete is imported (far majority from Mozambique). The production takes place mainly in one place per province. Only 0.7 million tons get exported (mainly Botswana and Namibia).

Obtained information:

Afrisam provided a detailed overview of their logistical operations for the last 5 years. The data from 2010 are the most recent and therefore used for our analysis. The origins are on a town/city level while the destinations are grouped by province. Therefore we took the middle points of the provinces to derive the transport distances.

Lafarge was unfortunately not in the position to provide actual figures but instead gave rough indications regarding their transport. These indications gave a comparable idea of the transportation in the cement business as obtained from Afrisam.

**Paper & paper products**

Paper and paper products can be divided in four categories; paper pulp, paper on rolls, A4 paper and paper tissues. The South African paper production market (4.4 million tons) is dominated by two companies Sappi (45%) and Mondi. Another 0.8 million tons is being imported (mainly through Durban). South Africa exports 0.9 million tons (also mainly through Durban).

Mondi has two big paper factories (Richards Bay and Durban) and Sappi has his plants in Johannesburg, Durban and Port Elisabeth.

Obtained information:

Sappi provided us with year round volume figures on a MD level and an average distance from their (not shared and therefore unknown) production plants to these regions.

**Sugarcane**

In total 20 million tons of sugarcane are produced in South Africa by 35.000 growers, all producing in the KwaZulu – Natal province. In total 14 sugar mills are operating in the same province. Because of economic feasibility the sugarcane only travels a maximum of about 60km on road between the growers and the mills.

The harvesting groups and transporting groups are responsible for all the transport of the sugar cane, which is done partly on rail but mainly on road.

All the growers are organized in so called local grower committees. These committees themselves are organized in the National Growers Association.

Obtained information:

The Sugarcane Association South Africa provided data on the total production, mode of transport and average distance from the growers to the mills for each of the fourteen mill regions.

Although no extra assumptions were necessary to process the data, the figures are averages and therefore show a very limited level of detail. Moreover the averages are clustered very close together so this makes it hard to make conclusions about the trip length distribution.

**Fruits (citrus, sub-tropical, deciduous, viticulture)**

Although the different types of fruit are grown in different areas their logistical behaviour is very alike. Fruits are used for export (about a third), another third gets sold on local markets and the last part is processed or canned.

The local markets get their fruit mostly from within their own province but some of it travels longer distances. The processors and canneries are traditionally based close to the fruit farms.

Citrus gets mainly exported through the seaports of Durban, Cape Town and Port Elisabeth. Deciduous fruits and grapes (viticulture) go through the port of Cape Town.

Obtained information:

Actual export flows for 2003 (for deciduous fruit we have 2010 export data). This information is confirmed to be still valid by the obtainers and providers of the information, Hortgro.

Information from the four biggest fresh produce markets about the origin of the fruits that are sold on their markets including volumes.

Information from a couple of fruit processors (Cape fruit processors) and canneries (RFF foods, Asthon canning, S. A. Preserving Co, Summerpride) regarding the distance from the fruit farms where they get their fruits from to their own facilities and the volumes that are being transported.

The export data are valuable as the specific origins, destinations and routes are provided. The provided information regarding the origin of the fruit from the fresh produce markets differed a lot from specific postal codes to province level. For the province level of detail we chose the distance between the market and the middle points of the fruit producing regions (within the indicated province). The food processors provided distance and volume figures.

All export figures, the majority of the local market information and part of the figures for the canning and processing industry were available. Because the distances in the canning and processing are limited it might be that the available data are slightly skewed to longer distances.

**Wheat**

The transportation in the wheat industry (about 2 – 3 Million tons) changed a few years ago when the industry shifted from a closely regulated market to an open market. In the closely regulated market all wheat was distributed to the closest silo and from there to the closest mill, mainly by train (80%). Since the opening of the market this percentage has decreased tremendously and more than 80% is now moved by road, often travelling far longer distances than necessary (from an ideal point of view). The wheat is now transported by (specialized) logistical companies like Grain Carriers, Afgri and Kaap Agri.

Obtained information

Kaap Agri, who is responsible for the majority of the wheat transportation in the Western Cape, provided us with year round data on the tons of wheat transported. These figures account for about 10-20% of the total wheat transportation in South Africa. The origins and destinations are given on a town/city level. Only distance needed to be calculated.

The provided data however show a noteworthy pattern. A lot of short distance low volume shipments (between 0 and 200km) nearly any shipments between 200 and 1000km and again many low volume shipments between 1000 and 1500km. This mainly has to do with the fact that Kaap Agri focuses on the farmers in the Eastern and Western Cape who have two major customers located in Cape Town (the 200km region) and Johannesburg (1000 -1500km region). This makes it challenging to use the data.

**Cotton**

The cotton market (including seeds, lint and yarn) is declining steeply in the last few years. In 2009 the total production of cotton was about 0.7 million tons of which a part has been imported. The raw cotton is transported from the growers to the ginners where the seeds get divided from the lint. The seeds are processed into oil or animal food. The lint is going to spinners who process it into yarn. The yarn is finally transported to the textile manufacturers. The used figures for cotton are based on the raw cotton only. Therefore the ginners are the points of demand.

Obtained information:

Information is obtained from the Cotton board in South Africa. They provided figures regarding the production of cotton in the different regions in South Africa and indicated the ginners in these regions. Verification by the ginners confirmed that the cotton is ginned at the closest ginners. The distance and volume could be derived.

**Jet fuel**

Jet fuel is produced by the same companies responsible for the regular fuel. The mode of transport however differs. Some airports (OR Tambo, the biggest airport, among others) get jet fuel delivered by pipe line but the majority of the airports are supplied by road. The small airports normally have only very few fuel supply as all airlines fill up at the (cheaper) bigger airports.

ACSA owns and operates 9 of the biggest airports in South Africa. They work closely together with the Department of Energy who keeps track of all the jet fuel consumption on the different airports operating in South Africa.

Obtained information:

One of the senior people within Shell SA provided us with the demand figures for the bigger airports and confirmed the assumption that they get supplied from the closest jet fuel producing refineries to minimize costs.

**Tobacco products**

The commodity of tobacco products exist of cigarettes, cigars and pipe tobacco. The cigars and pipe tobacco are insignificant compared to the volumes of cigarettes which are produced by one major player, Britisch American Tobacco (89%) and two smaller ones JIT (6%) and PMI (5%).

The total consumption of tobacco products in South Africa is about 662.000 tons. Very little gets imported but the export through the port of Durban is significant (however exact figure are not available).

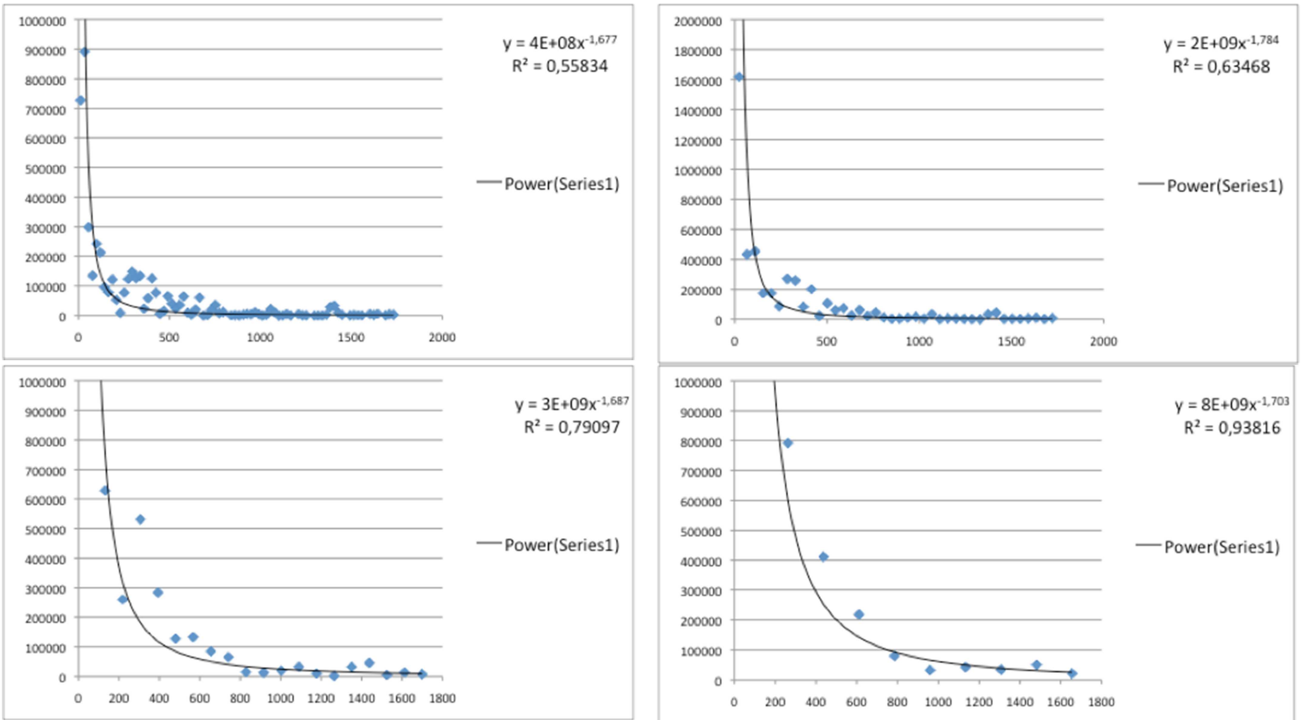
BAT has only one big factory in Heidelberg (Gauteng) and supplies from there the whole of South Africa.

Obtained information:
Unfortunately none of the tobacco product manufacturers were able to provide us with the requested data. But the fact that BAT is responsible for about 90% of the total supply of the South African market and very few of their products get imported we can, based on the demand figures of Conningarth derive the trip distribution (volume vs. distance) by supplying all demand points from the Heidelberg (Gauteng) factory of BAT.

## Appendix H Influence of binning on decay parameter value

By the nature of the procedure one expects binning to influence the decay parameter value. In the following example it is shown that the binning procedure has a certain influence on the decay parameter value. However this influence is not the same for all commodity groupings.

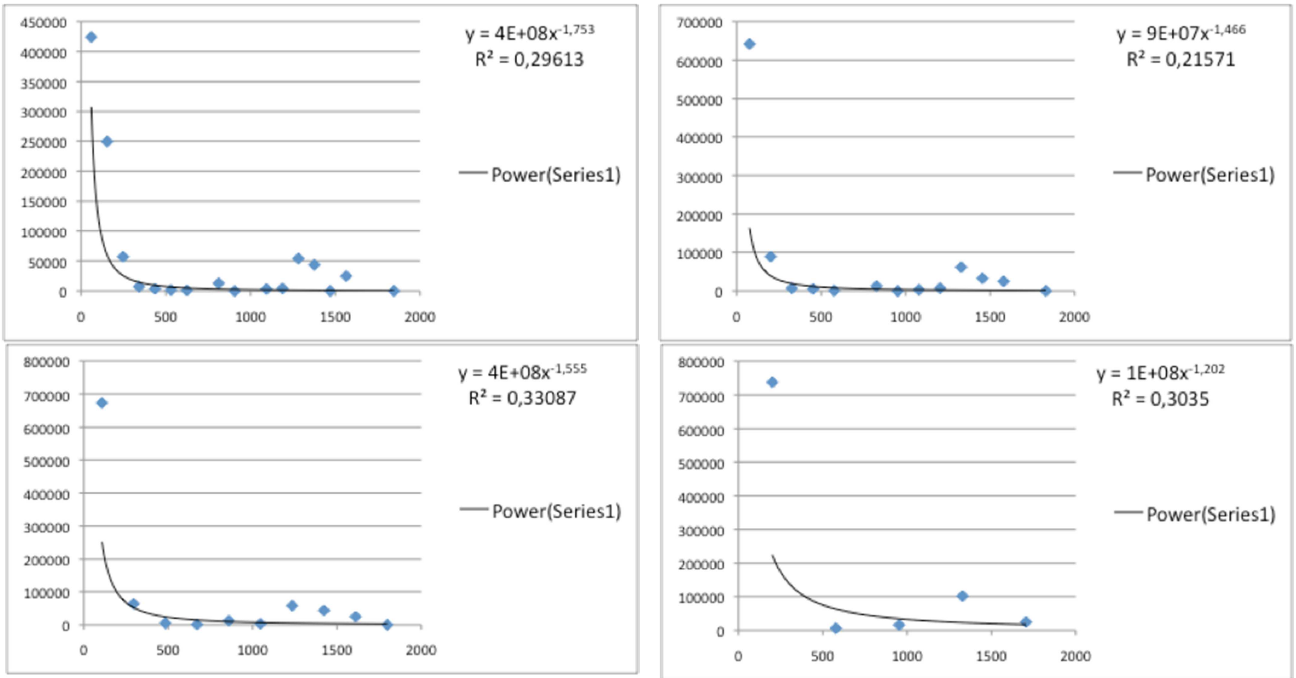
We will show this by using the commodity groupings beverages and deciduous fruits as examples.




Figures H.1 – 4 Curve fitting through observations classified in different bin sizes (Beverages)

| Amount of bins applied | Decay parameter value |
|---|---|
| 80 | 1.677 |
| 40 | 1.784 |
| 20 | 1.687 |
| 10 | 1.703 |

Table H.1 Relation between bin size and decay parameter value (Beverages)




Figures H.5 – 8 Curve fitting through observations classified in different bin sizes (Deciduous fruits)

| Amount of bins applied | Decay parameter value |
|---|---|
| 40 | 1.753 |
| 20 | 1.466 |
| 10 | 1.555 |
| 5 | 1.202 |

Table H.2 Relation between bin size and decay parameter value (Deciduous fruit)

Although the bin sizes in Figure H1 – 4 have been doubled between the successive graphs the decay parameter for beverages has only slightly changed. The decay parameter for deciduous fruits has clearly changed more than the one for beverages. Therefore the choice of the amount of bins/ bin size seems to have more impact for deciduous fruits. However if we take robustness regarding the mean trip length (MTL) into account (Section 5.8) we see that a change in the decay parameter value from 1.75 to 1.2 implies only an increase in MTL of 20 km. Therefore the bin size does not seem to have a significant influence in the application of the gravity model. It may however influence the correlation analysis in Chapter 6.

The difference between the influences of the binning procedure has probably not so much to do with the commodity grouping at hand but with the amount of available observations. In case of limited amount of observations changing the bin size will have much more impact (bins go for example from zero to three observations by a small change in bin sizes) than when many (well spread) observations are available (like in the case of beverages).

## Appendix I Overview of new derived decay functions

| **Commodity grouping** | **Decay function** | **Decay parameter** | **R^2** | **ATD** | **Alternative decay parameter*** | **R^2** | **Comments** |
|---|---|---|---|---|---|---|---|
| Beverages | Power | 1.68 | 0.56 | 286km | 0.0032 | 0,14 | 80 bins (74 filled with observations) Only primary distribution (brewery to depot) |
| Bricks | Exponential | 0.01 | 0.56 | 120km | 0.61 | 0,33 | Only bricks (already binned data) |
| Cement | Exponential | 0.0068 | 0.37 | 310km | 1.71 | 0,27 | No bins |
| Fruit<br>*Deciduous fruit | Power | 1.58 | 0.33 | 723km | 0.002 | 0,16 | 20 bins (16 filled with observations) |
| *Citrus | Exponential | 0.0022 | 0.28 | 608km | 1.32 | 0,023 | 20 ins |
| *Sub-tropical | Power | 1.088 | 0.15 | 625km | 0.0008 | 0,00 | 10 bins |
| *Viticulture | Exponential | 0.0034 | 0.29 | 263km | 1.89 | 0,069 | 20 bins (16 filled with observations) |
| Jet fuel | Power | 1.63 | 0.53 | 202km | 0.0054 | 0,27 | No bins |
| Motor vehicles | Power | 0.040 | 0.00 | 607km | 0.0001 | 0,00 | No bins |
| Paper & paper products | Exponential | 0.0021 | 0.41 | 443km | 1.11 | 0,018 | 10 bins (9 filled with observations) |
| Sugar Cane | Exponential | 0.052 | | 30.7km | 1.72 | | Based on a calculated ATD of 30km |
| Wheat | Power | 0.58 | 0.30 | 553km | 0.0009 | 0,036 | 80 bins (39 filled with observations) |
| Cotton | Power | 0.75 | 0.54 | 664km | 0.006 | 0,47 | No bins |
| Tobacco | Power | 1.1 | 0.26 | 425km | 0.0022 | 0,067 | 40 bins |

Table I.1 Overview of new derived decay functions

* In case of a power function the alternative is an exponential function and vice versa

## Appendix J Correlation analyses based on new derived decay functions

In the scatter plots and tables below the outcome of the correlation analyses between decay and possible influencing factors is presented. The first correlation analysis is, unlike the analysis in Chapter 6, based on only 8 commodity groupings. The negative power decay function was best fitted to the gathered data for these commodity groupings. The second correlation analysis is based on the 6 commodity groupings for which the negative exponential function fitted best to the obtained data. The last correlation analysis shows the relation between decay and the possible influencing factors for all 14 commodity groupings when they would have been assigned a negative exponential function. For the first two correlation analyses it is important to keep in mind that the amount of observations is actually too low to really derive conclusions from the outcome. We only present the scatter plots of relations that deserve extra attention based on the outcome of the correlation analyses.

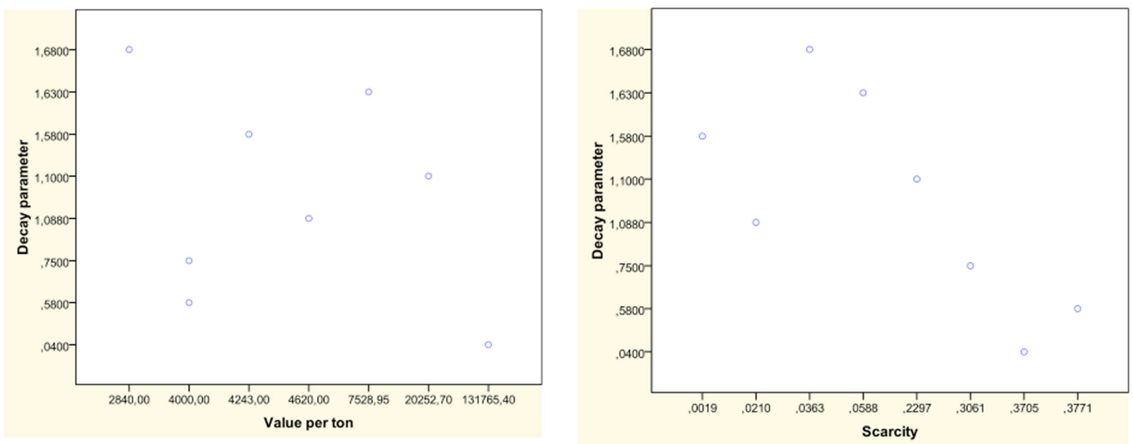



Figure J.1 Visualization of the correlation between decay and value per ton (left) and decay and scarcity (right)*

| | | Value per ton | Scarcity | Supply concentration | Local demand | Homogeneity | Brand |
|---|---|---|---|---|---|---|---|
| Decay parameter | Pearson Correlation | -,707 | -,880 | -,027 | ,565 | -,613 | -,038 |
| | Sig. (1-tailed) | ,025 | ,002 | ,474 | ,072 | ,053 | ,464 |
| | N | 8 | 8 | 8 | 8 | 8 | 8 |

Table J.1 Correlation coefficient table of the correlation between decay and the possible influencing factors *

* Based on 8 commodity groupings with a negative power decay function as best fit to data

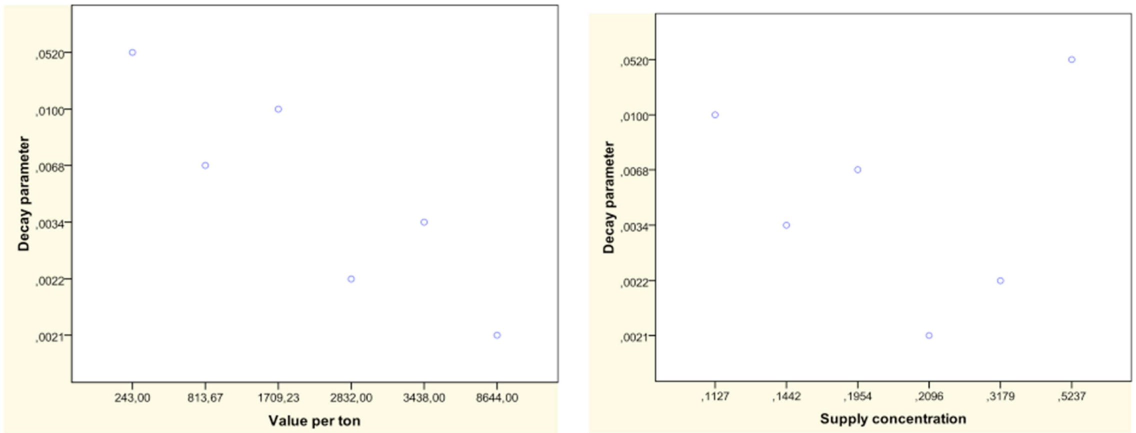


Figure J.2 Visualization of the correlation between decay and value per ton (left) and decay and supply concentration right) **

| | | Value per ton | Scarcity | Supply concentration | Local demand | Homogeneity | Brand |
|---|---|---|---|---|---|---|---|
| Decay parameter | Pearson Correlation | -,521 | -,327 | ,826 | -,415 | -,652 | -,503 |
| | Sig. (1-tailed) | ,144 | ,263 | ,021 | ,207 | ,080 | ,155 |
| | N | 6 | 6 | 6 | 6 | 6 | 6 |

Table J.2 Correlation coefficient table of the correlation between decay and the possible influencing factors **

** Based on 6 commodity groupings with a negative exponential decay function as best fit to data

| | | Value per ton | Scarcity | Supply concentration | Local demand | Homogeneity | Brand |
|---|---|---|---|---|---|---|---|
| Decay parameter | Pearson Correlation | -,195 | -,293 | ,347 | -,236 | -,086 | -,289 |
| | Sig. (1-tailed) | ,253 | ,155 | ,112 | ,208 | ,385 | ,158 |
| | N | 14 | 14 | 14 | 14 | 14 | 14 |

Table J.3 Correlation coefficient table of the correlation between decay and the possible influencing factors ***

*** Based on 14 commodity groupings. all assigned a negative exponential decay function

The correlation coefficient tables show some significant relations of moderate strength. However the squatter plots do not all visually underline these relations. In Figure J.2 the relationship between decay and value per ton seems present but this is not significant given the correlation coefficient table (Table J.2).

# Appendix K Regression model based vs. current decay parameters

| | | Predicted decay parameter* | | Current decay |
|---|---|---|---|---|
| Commodity group | Scarcity | With 'Bricks' | Without 'Bricks' | parameter* |
| BARLEY | 0.35 | 0.61 | 0.58 | 0.50 |
| COTTON | 0.31 | 0.72 | 0.72 | 0.15 |
| DECIDUOUS FRUIT | 0.00 | 1.52 | 1.65 | 0.15 |
| CITRUS | 0.00 | 1.53 | 1.65 | 0.15 |
| SUBTROPICAL FRUIT | 0.02 | 1.47 | 1.59 | 0.15 |
| VITICULTURE | 0.00 | 1.53 | 1.66 | 0.15 |
| GRAIN SORGHUM | 0.14 | 1.16 | 1.22 | 0.50 |
| LIVESTOCK (SLAUGHTERED) | 0.10 | 1.26 | 1.34 | 0.01 |
| MAIZE | 0.00 | 1.52 | 1.65 | 4.00 |
| SOYA BEANS | 0.19 | 1.03 | 1.07 | 0.50 |
| SUNFLOWER SEED | 0.09 | 1.29 | 1.38 | 0.50 |
| VEGETABLES | 0.02 | 1.47 | 1.59 | 0.01 |
| WHEAT | 0.38 | 0.54 | 0.50 | 0.50 |
| POULTRY PRODUCTS | 0.00 | 1.53 | 1.65 | 0.01 |
| DAIRY | 0.01 | 1.50 | 1.62 | 0.01 |
| SUGAR CANE | 0.00 | 1.53 | 1.66 | 0.75 |
| OTHER AGRICULTURE | 0.29 | 0.76 | 0.76 | 0.50 |
| COAL MINING | 0.01 | 1.51 | 1.64 | 0.64 |
| CRUDE PETROLEUM & NATURAL GAS | 0.91 | -0.87 | -1.14 | 0.42 |
| IRON ORE (HEMATITE) | 0.00 | 1.53 | 1.65 | 2.78 |
| MAGNETITE | 0.00 | 1.53 | 1.65 | 0.50 |
| CHROME | 0.00 | 1.52 | 1.64 | 2.04 |
| COPPER | 0.84 | -0.67 | -0.91 | 0.50 |
| MANGANESE | 0.00 | 1.53 | 1.65 | 0.13 |
| TITANIUM | 0.00 | 1.52 | 1.64 | 0.50 |
| ZINC | 0.64 | -0.16 | -0.31 | 0.50 |
| OTHER NON-FERROUS METAL MINING | 0.66 | -0.22 | -0.38 | 0.50 |
| STONE QUARRYING, CLAY & SAND-PITS: GRANITE | 0.07 | 1.35 | 1.45 | 0.60 |
| STONE QUARRYING, CLAY & SAND-PITS: LIMESTONE & LIME WORKS | 0.00 | 1.52 | 1.65 | 0.60 |
| STONE QUARRYING, CLAY & SAND-PITS: OTHER | 0.00 | 1.52 | 1.65 | 0.60 |
| MINING OF CHEMICAL & FERTILIZER MINERALS | 0.18 | 1.06 | 1.11 | 0.60 |
| OTHER NON-METALLIC MINERALS | 0.15 | 1.14 | 1.21 | 0.50 |
| OTHER MINING | 0.00 | 1.53 | 1.65 | 0.50 |
| FOOD AND FOOD PROCESSING | 0.09 | 1.29 | 1.37 | 0.15 |
| BEVERAGES | 0.04 | 1.43 | 1.54 | 0.15 |
| TOBACCO PRODUCTS | 0.23 | 0.92 | 0.95 | 0.15 |
| TEXTILES, CLOTHING, LEATHER PRODUCTS AND FOOTWEAR | 0.46 | 0.31 | 0.23 | 0.15 |
| WOOD AND WOOD PRODUCTS | 0.08 | 1.31 | 1.40 | 0.50 |
| FURNITURE | 0.24 | 0.90 | 0.93 | 0.15 |
| PAPER & PAPER PRODUCTS | 0.15 | 1.13 | 1.20 | 0.15 |
| PRINTING AND PUBLISHING | 0.08 | 1.33 | 1.42 | 0.01 |
| INDUSTRIAL CHEMICALS | 0.28 | 0.80 | 0.81 | 0.15 |
| FERTILIZERS AND PESTICIDES | 0.29 | 0.77 | 0.77 | 2.00 |
| PHARMACEUTICAL, DETERGENTS AND TOILETRIES | 0.41 | 0.46 | 0.41 | 0.15 |
| PETROLEUM REFINERIES AND PRODUCTS OF PETROLEUM/COAL | 0.17 | 1.07 | 1.12 | 2.00 |
| RUBBER PRODUCTS | 0.57 | 0.02 | -0.10 | 0.15 |
| OTHER CHEMICALS | 0.12 | 1.22 | 1.30 | 0.15 |
| NON-METALLIC MINERAL PRODUCTS | 0.11 | 1.24 | 1.32 | 0.20 |
| BRICKS | 0.01 | 1.50 | 1.63 | 0.58 |
| CEMENT | 0.18 | 1.04 | 1.09 | 0.36 |
| FERROCHROME | 0.01 | 1.51 | 1.63 | 0.01 |
| FERROMANGANESE | 0.00 | 1.52 | 1.65 | 0.01 |
| OTHER IRON AND STEEL BASIC INDUSTRIES | 0.18 | 1.04 | 1.09 | 0.60 |
| NON-FERROUS METAL BASIC INDUSTRIES | 0.06 | 1.37 | 1.47 | 2.00 |
| METAL PRODUCTS EXCLUDING MACHINERY | 0.14 | 1.17 | 1.24 | 0.08 |
| MACHINERY AND EQUIPMENT | 0.47 | 0.28 | 0.20 | 0.15 |
| ELECTRICAL MACHINERY | 0.50 | 0.22 | 0.14 | 0.15 |
| MOTOR VEHICLES | 0.37 | 0.55 | 0.52 | 0.01 |
| MOTOR VEHICLE PARTS AND ACCESSORIES | 0.49 | 0.24 | 0.15 | 0.01 |
| TRANSPORT EQUIPMENT | 0.15 | 1.12 | 1.18 | 0.20 |
| OTHER MANUFACTURING INDUSTRIES | 0.35 | 0.60 | 0.57 | 0.15 |
| WATER SUPPLY | 0.02 | 1.49 | 1.61 | 0.15 |
| GAS | 0.01 | 1.51 | 1.63 | 2.00 |
| JET FUEL | 0.06 | 1.37 | 1.47 | 2.00 |

Table K.1 Predicted decay parameters based on regression (including and excluding ‘Bricks’) vs. currently applied decay parameters

* All decay parameters are based on negative power functions