DEFICIENT TESTING DATABASES: A RELIABILITY-DRIVEN EVALUATION OF PRIVACY MODELS PROVING A TRADE-OFF BETWEEN DATA INTEGRITY AND RE-IDENTIFICATION RISK

Florian Ventur

Industrial Engineering & Management

June 2020

topicus

UNIVERSITY OF TWENTE.

This report is commissioned by Topicus Overheid BV and executed in the context of the bachelor program Industrial Engineering & Management at the University of Twente.

Topicus Overheid BV

Postbus 317

7411 HW Deventer

Tel. +31 570 662 662

University of Twente

BSc Industrial Engineering & Management

Postbus 217

7500 AE Enschede

Tel. +31 534 89 91 11

Deficient testing databases: A reliability-driven evaluation of privacy models providing a trade-off between data integrity and re-identification risk.

Bachelor Thesis Industrial Engineering & Management

Author

Florian Ventur

s1859862

f.ventur@student.utwente.nl

Supervisor Topicus Overheid

BSc Y. Koen (Yoshi)

1st Supervisor University of Twente

dr. A. Abhishta (Abhishta)

Faculty of Behavioural, Management and Social Sciences

Supervisor Topicus Overheid

BSc D. Verbeek (Dirk)

2nd Supervisor University of Twente

dr.ir. J. Goseling (Jasper)

Faculty of Electrical Engineering, Mathematics and Computer Science

Acknowledgements

The completion of this report was only possible with the involvement of certain people. Hereby, I would like to express my appreciation to those who guided me with their expertise through this academic year. I acknowledge all your feedback and especially your approachability.

Dr. Abhishta Abhishta

Dr. ir. Jasper Goseling

Yoshi Koen

Dirk Verbeek

Staff of Topicus Overheid BV

Foreword

Dit verslag markeert de eindstreep van mijn tijd in Enschede. Met plezier en weemoed kijk ik terug op vier gevarieerde jaren waarin ik me kon ontplooien op het vlak van continuïteit en spontaniteit. Hoewel dit A4'tje nooit genoeg zal zijn om elke bijdrage expliciet te waarderen, volgt alsnog een poging.

Allereerst wil ik mijn ouders bedanken die mij telkens de vrijheid gaven om elke weg in te slaan en me daarin ook te ondersteunen, hetzij direct, hetzij indirect. De opvrolijkpakketjes in stressvolle fases of de adviezen over mijn studieplanning hebben me altijd verder geholpen.

Vanaf dag één in de Kick-in werd ik, mede door een administratieve fout, warm opgenomen door mijn huidige doegroep V.A. Coq. Zonder hun begrip voor mijn beperkte woordenschat had ik me amper zo goed kunnen integreren in het Enschedese studentenleven. Vooral dank aan quarantaine buddy en eerste podcastproducent Pim, lapplandavonturier en tweede podcastproducent Matthijs en blij ei Wieneke.

Daarnaast heeft de D.R.V. Euros en de ware Beukende Bokken me door de gespierdste en vormlooste fases van mijn studententijd geholpen. Nog steeds ben ik trots dat we op het Amsterdam-Rijnkanaal als 11e met Aeolus finishten en talloze mooie momenten buiten en binnen de boot hebben beleefd. Ik kijk nu al uit naar de volgende gelegenheid om Biddinghuizen een bezoekje te brengen.

Verder ben ik fier om deel uit te maken van Solar Boat Twente. Hoewel mijn studieachtergrond – zacht uitgedrukt – niet perfect aansloot op mijn werkzaamheden, kon ik toch mijn technische kennis flink verbreden. Mijn grootste dank aan het hele team die met mij het jaar hebben beleefd en de Twentsch Pegasus op de kaart hebben gezet. De vroege ochtendritjes met Jurriën, de heugelijke avonden aan de Irisstraat en vooral de leuke activiteiten met het hele team hebben me een waardevol jaar opgeleverd.

Ook mijn afstudeerbedrijf Topicus heeft hieraan bijgedragen. Elke keer als de deadlines van Solar Boat aankwamen, werd begrip getoond en mocht ik mijn tijd flexibel indelen. De fijne werksfeer en communicatie hebben de, in mijn ogen, optimale werkvloer aan het licht gebracht. Ik wil Dirk bedanken voor zijn enthousiasme over mijn ideeën en inbreng, Yoshi voor de geweldige begeleiding en het hele Team Legends voor jullie talloze uitleggen van ICT-gerelateerde onderwerpen.

Tot slot dank aan de onverwoestbare SoWi Elite aan die ik al sinds jaren gehecht ben, Huize Barbapapa voor de gezellige tijd, wereldverbeteraar & filosoof Klaske, levensgenieter Mello en mister binnenvaartschipper Sven.

Executive Summary

Managers and administrators of data-rich environments take on responsibility in the protection of personal data. In the previous five years, data protection laws as the General Data Protection Regulation (GDPR) or the California Consumer Privacy Act (CCPA) were implemented and enforced to guarantee conscientious data handling. This research aims at the investigation of data transformation by means of anonymization for secure data storage, processing and handling. In addition to previous research, the trade-off privacy versus data integrity is expanded by the concept of data diversity. The two conflicting variables are weighted by the harmonic mean 'F-score' which is widely used in the evaluation of algorithms and systems. Privacy is operationalized by the re-identification threat which indicates the probability of still identifying an individual after anonymization. Data integrity is defined in the further course.

The research question reads as follows:

"Which privacy model provides a harmonic mean of data integrity and re-identification threat aiming the reduction of bug frequencies in software testing by the preservation of realistic and diverse data?"

As software developers ask for realistic and representative anonymization, data integrity is lighted from two perspectives. Firstly, the preservation of an entities' context which allows a value to be generalized from e.g. 'Linux' to 'operating system' without changing its information integrity. Secondly, the preservation of an entities' data property which allows a value to be overwritten from e.g. 'Linux' to 'Windows' which is not correct but remains data diversity. Both scenarios are evaluated regarding the usability for Business Analysts (integrity driven) and software developers (property integrity and diversity driven).

The data used for the analysis is extracted from the production database of the municipality of Enschede. Several data tables are joined to three representative datasets containing personal data and sensitive attributes as problem categories or job functions. In total, three anonymization methods (also called privacy models) are evaluated against each other.

The analytical results are two-folded. In the case of attribute homogeneity (e.g. residence in municipality of Enschede), the privacy model ***l*-diversity** provide good results in terms of privacy and utility. Parameters can be set relatively strict ($l = 4$) as the class size of a homogenous dataset is per definition bigger. Next to that, the **average re-identification risk model** with 5% provides equally good results but with a bigger re-identification risk spread on average. In the case of attribute diversity (e.g. ZIP codes), a diversity is also convenient but particularly with less strict parameters ($l = 2$). For the transformation methods, generalization is seen as the preference for Business Analysts (good data integrity) and microaggregation is seen as the preference for software developers (good data diversity).

The research shows that all well-performing transformed datasets, in comparison to the reference production datasets, unaltered in size and different in representativeness and diversity. In each scenario, an authentic size of a dataset can approach realistic response times which solves a part of the bugs. Data readability, integrity & diversity are dependent on, logically, the privacy model and the re-identification treat level which unitedly solve the major bugs as falsified query times, display errors or wrong indexing. In contrast to the existing testing database, those measurement variables are surpassed for most of the proposed privacy models. Thereby, the desired data properties are met to a degree which is determined by the re-identification treat level. By means of a full implementation, the concerned employees of Topicus can reduce a fraction of the maintenance workload (10 to 25% of total workload) which is dedicated to bug fixing.

Several recommendations can be given. Regarding future implementation of privacy models to data environments, it is advisable to create or acquire detailed reference data tables for creating generalization hierarchies. This is already done for the used data tables in this research, but more extensive and detailed hierarchies can improve data integrity and data diversity significantly.

Next to that, users tend to enter sensitive personal information in input fields for e.g. a new alert or in the problem description. This complicates the anonymization as BSN's or telephone numbers are processed in a text and is, naturally, not necessary as the text field is already coupled to a data table where the personal information is listed. A small user note in Gidso might prevent this unnecessary privacy risk. Besides, drop-down menus for job functions or companies standardizes input data and improves the quality of anonymization subsequently.

Reader's Guide

This report identifies possible solutions of data transformation by means of anonymization methods aimed to reduce the occurrence of bugs caused by deficient testing environments. The content of the covered chapters is listed below.

Chapter 1 provides the context of the research. Concretely, the relevance of anonymization, the background of Topicus and Gidso, the problem statement and scope & limitations are stated.

Chapter 2 addresses the theoretical framework. Essential knowledge about anonymization, external validity and data management are covered.

Chapter 3 contrasts the existing data environment with the desired data environment in the context of data properties.

Chapter 4 specifies the execution of data management, the selection of privacy models and the related measurement variables.

Chapter 5 evaluates the selected privacy models employing pre-defined measurement variables.

Chapter 6 qualifies the measurement outcomes and provides a recommendation of possible privacy models aiming the reduction of bug occurrences conditioned by privacy thresholds and desired data attributes.

TABLE OF CONTENTS

List of Figures

List of Tables

Abbreviations

ARX	ARX Data Anonymization Tool
BSN	Burgerservicenummer (EN: Citizen service number)
BV	Besloten vennootschap (EN: Private limited company)
CBS	Centraal Bureau voor de Statistiek (EN: Central Agency for Statistics)
DSRM	Design Science Research Methodology
ER-Model	Entity–Relationship Model
CCPA	California Consumer Privacy Act
GBA	Gemeentelijke basisadministratie voor persoonsgegevens (EN: Municipal Personal Records Database)
GDPR	General Data Protection Regulation
ID	Identifier
QI	Quasi-Identifier
SaaS	Software as a Service
SLA	Service-Level-Agreement
SQL	Structured Query Language
WMO	Wet maatschappelijke ondersteuning (EN: Social Support Act)
ZIP-code	Zone Improvement Plan

1 Introduction

This chapter addresses the relevance of anonymization in 1.1 Background, the company and its product in Context 1.2, the structural obstacles of inadequate testing environments in problem statement 1.3 and the problem-solving approach in 1.4 Scope & Limitations.

1.1 Background

The European Union drafted and passed the General Data Protection Regulation (GDPR) in 2016 which is considered as the toughest privacy and security laws in the world (Satariano, 2018). Since it covers the protection of all EU citizens' personal data, data environments where health-related information is stored and processed should be treated with extra security. In the case of working frequently with sensitive data, pseudonymization or anonymization should be considered to avoid privacy violation as regards data leakage. In any case, unauthorized persons should never be able to interpret that data to valuable information.

1.2 Context

1.2.1 Topicus

Topicus is a leading IT software provider on the Dutch market, offering impactful and user-friendly software solutions in the world of finance, education, healthcare and social services. Their core business revolves around administrative process management, Software as a Service (SaaS) and integration while aiming to assist clients in remodeling and improving their service to customers to add value to humankind and society. On an average, every Dutch citizen is supported 2.3 times by all Topicus software combined directly or indirectly (Topicus, 2020). This extends from administrative software in primary school to automated business financing platforms.

1.2.2 Gidso

Topicus Overheid (Social Services) developed their flagship platform called Gidso. It serves as coupling point for, among others, triage & directing capped by declaration processing and effect measurements. It provides the tools to connect municipalities, clients and third parties by providing an application which automates administrative processes and simplifies communication lines. This automatization process does not only happen internally but also with external software providers via busses or the standardized Municipal Personal Records Database (in Dutch: GBA). Gidso's users are municipalities which serve as a contact point for the citizens. These citizens make use of the Social Support Act (in Dutch: WMO) and ask the municipalities for

help related to physical disabilities, home care or psychological issues. A record about the client, his relations and a possible roadmap for treatments is then drawn up within Gidso and potentially third parties as doctors or healthcare suppliers are connected to the record.

1.3 Problem Statement

As Gidso is deployed as a SaaS, the ownership of personal data is fully assigned to the municipalities. Accessibility of realistic data is limited by several national and European data protection laws. This restriction is circumvented by transformed or synthetic datasets. As a result, companies elude this barrier of re-identification while undergoing serval testing issues as described further.

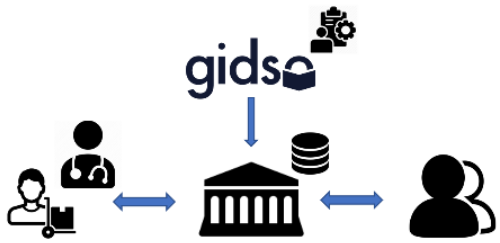


Figure 1.1 | Relationship between Software Provider and User

Topicus takes privacy seriously. One of the main threats regarding data leakages are incautious employees which are victims of data thefts. Therefore, real sensitive data should be only accessible by trusted administrators. This is convenient in terms of security but disadvantageous when it comes to testing the performance of a just build application extension or improvement against possible real-life scenarios. Over time, the software developers created some own fictional client records with either specific properties for their testing purposes or some nonsensical records to fill their database fast. Beside to this testing database, the software developers can also access a semi-realistic database (acceptance database) under certain conditions which are held and used by the municipalities to test the release versions from Topicus. However, this is only done in exceptional cases since it is forbidden using foreign data except if data is purely used for software improvements. Operational testing procedures are not covered by this exception. This results in being forced to use the own non-realistic testing database sets which can hardly be used to investigate outlier cases. Outlier cases can be categorized in excessive query times, display errors, inconvenient indexing or further unknown errors. Those bugs limit the usability by the customer and higher the uncertainty about meeting the set Service Level Agreements (SLA's). The current SLA is set at 98.5% where the application should respond within the arranged query times.

At this moment, even the most recent database is characterized by repetitive and data-poor client records which interfere with the SQL service buffer tool and database indices. Although the buffer functions as a timesaving component for repetitive queries within the testing phase, it does not take more realistic and diverse queries into account which are called from the hard disk on the SQL server. This is a reinforcing

factor for the unrealistic query time which arises due to the inconvenient number of 150 personal records in the testing environment.

In addition to the lowered usability at the customer side, the company facing complaints from the municipality which should be fixed. Although the software developer divisions use the method Kanban for structurally fixing errors, unnecessary bugs hinder the speed of further software development. Therefore, complaints which are directly connected to invalid testing should be canceled out to both save time on one task and increase progress flows due to less software development iteration cycles. A problem cluster stating causes and consequences is developed and illustrated in Figure 1.2. for identifying the action problem analogous to the method of H. Heerkens (2017).

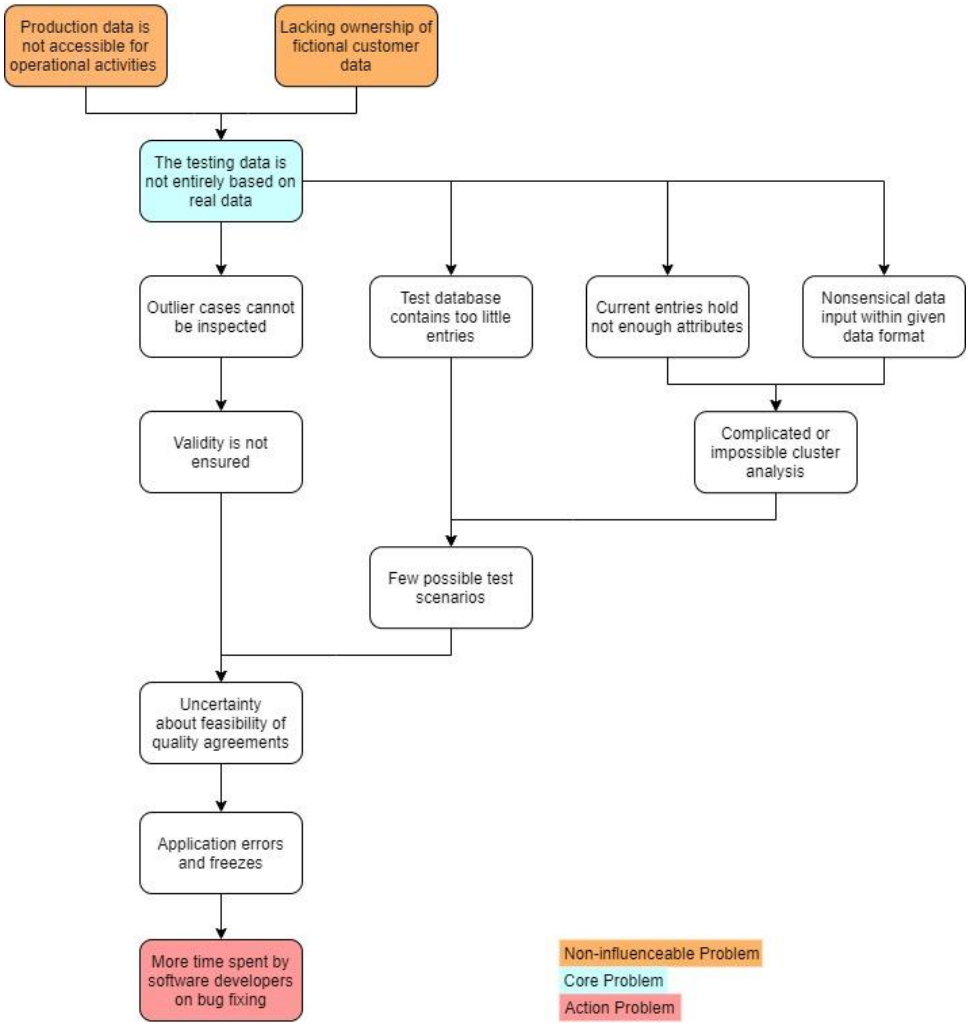


Figure 1.2 | Problem Cluster

1.4 Scope & Limitations

By breaking down the problem statement, the global desired improvements get clear. The current database does not meet the required requirements for testing purposes and two possible solving approaches are viable. Either generating data from scratch based on roughly estimated input variables or transform real personal data such individuals cannot be backtracked. Since great inaccuracy occurs when implementing the first solution, a consensual favor is given to the anonymization.

Topicus currently uses PostgreSQL as their relational database management system. Although the notation of SQL is standardized, PostgreSQL let users write functions in a wide variety of object-orientated programming languages. This is the only hard constraint and it is advisable to continue working in this system. Next to that, the artifacts of this research do not intersect with the existing software used by the developers, so the selection of tools is completely footloose and fancy-free. This relieves the realization of the pre-defined scope since selection is adaptable to complexity and scale. For the sanitization itself, an established anonymization software is evaluated against factors like usability, cost and popularity.

1.4.1 Research Objective

The goal of this research is to evaluate privacy models based on the properties of existing production data tables stored by Topicus. As datasets are just a subset of a whole database, the data table selection aspires to be representative and data rich. Central to the analysis is the anonymization applicability to the remaining data tables in the data environment as no all-compounding implementation can be given in the given time frame. However, the report and sanitized data tables are aimed to give a detailed roadmap on the preparation and transformation of data. The final anonymization iterations are presented from a statistical and a textual point of view for guaranteeing maximal transparency.

Additionally, this research is accompanied by advice about the realization of data transformation on large scale stimulating the ease and rapidity of anonymization. Topicus professionals guided in that way that immediate data migration between local servers can take place straightforward without needed detailed theoretical knowledge about anonymization. By this. the action problem of unnecessary bug fixing is effectively solved as theory flows into application and connects simultaneously to the expertise of the Topicus software engineers or Data Scientists.

1.4.2 Research Question

Based on the action problem and the GDPR constrain, the central research question can be derived addressing the conflicting factors of privacy and utility in the context of bug frequency reduction.

> Which privacy model provides a harmonic mean of data integrity and re-identification threat aiming the reduction of bug frequencies in software testing by the preservation of realistic and diverse data?

1.4.3 Methodological framework

The preference for a framework is given to the Design Science Research Methodology (Peffers, K., et al, 2007). This methodology is popular in the field of Information Systems and is directed to knowledge acquisition of configuration embodiment, structure, composition, purpose, value and meaning in man made things and systems. As the data transformation is guided by a rigid IT-infrastructure (ER framework) and several conditions on shaping the anonymization, the methodology fit is comprehensible. Also, the transformed datasets are, in fact, a man-made product.

Additionally, the DSRM includes a detailed description of the Design & Development phase which is especially in IT projects a time-consuming phase due to numerous iterations (Rodriguez-Repiso, L., et al, 2007). By this, the research is likely to be less error prone as the phase specification is little interpretable.

A detailed description of every phase can be found in the Appendix.

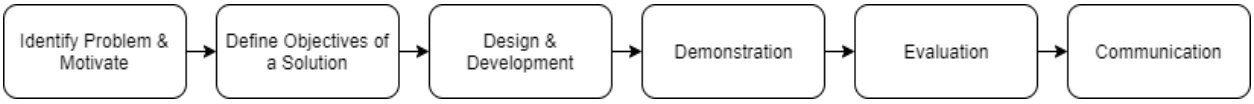


Figure 1.3 | Framework Design Science Research Methodology (DSRM)

While some phases seem self-explanatory, it is essential to point out the link between each stage. The research is namely purely based on this framework for providing a continuous logical coherence. The theoretical and practical aspects of each phase are addressed in chapter 1.4.4.

1.4.4 Plan

Each central research question is guided by the answer of several sub-questions. Unitedly, they also serve as a roadmap for the answer of the research question as reasoning and arguments are derived by the sub-questions. In the following, the phases of the DSRM are explained and concretized with relevant questions for each chapter.

The *first phase* (chapter 1.3) is initiated by specifying the research problem and conceptually delineate the given issue. A systematic cause-effect scheme is described by H.Heerkens (2017) which clearly puts all related problems into a perspective.

The *second phase* derives the research focus in terms of feasibility and limitations (chapter 1.4). Also, the type of research and the related work for research specification are stated which are addressed in the whole chapter 2. Concretely, the following sub-questions are set:

1. Which common anonymization models exist in recent literature?
    a. By which law or regulation is the sanitization guided?
    b. How do non-perturbative and synthetic data methods score, generally spoken, in terms of record and population integrity?
    c. How do non-perturbative and synthetic data methods score, generally spoken, in terms of record re-identification?
    d. Which methods are most resistant against linkage and probabilistic attacks?
    e. What contextual data is needed for non-perturbative and synthetic data methods?
2. Which methods are available for measuring representativeness and diversity of the data table selection?
3. How should personal records be prepared before running the anonymization iterations?

To get a sense of privacy, the legal framework is essential for the whole research. Basic and more advanced techniques are evaluated relative to each other as no absolute state of privacy exists without dropping all data. As measurements for universal implementations are highly dependent on representativeness regarding a reference population, a sub-chapter about external validity is included. Data preprocessing lowers the error rate of analysis results and therefore, round off chapter 2.

The *third phase* (chapter 3&4) evolves around the operationalization of key variables and proposed models which are intended to apply. Related to this research, sampling from testing database, diversity measurements and privacy as k-anonymity or microaggregation are riddled with parameters in this phase.

4. How do the properties of the test and production database differ?
    a. Which data tables take on a central role in the database environment?
    b. How does the data type distribution look regarding potential data distinguishability?
    c. What is the percentage of filled out input fields in the test & production database?
    d. How diverse and representative is the current database?
5. Which data sets are used for the anonymization?
    a. Based on which Quasi Identifiers (QI's) are sensitive terms detected?
    b. What contextual data is needed?
    c. How are the data records pre-processed?
    d. What privacy models are used?
    e. Which parameters are allocated to the privacy models?
    f. Along which output values is the analysis evaluated?

Addressing the disparity between the reality and the desired norm by comparing the testing with the production database is done by the sub-questions of question 4. Data management and parameter selection is subsequently the focus of question 5.

The *fourth phase* envisages the implementation of the pre-defined research model. By this, data in the form of measurements are set out and compared with each other.

6. How are, due to software restrictions, non-processable ID's transformed?
7. What are the results of the applied data transformations?
    a. What are the respective generalization intensities & Record-based errors?
    b. What are the respective re-identification risks?
    c. What are the respective average class sizes?
    d. What is the respective F-score?

Qualifying and Interpretation the data from the model into meaningful information is the *fifth phase* of the methodology and realized in section 7.1.

In the *sixth phase* in section 7.2 (et seqq), the results will be presented to relevant stakeholders as researches but most naturally the company direction. Recommendations for further research are also given. This is done in oral and in written form.

2 Theoretical framework

This chapter specifies techniques for data anonymization with ensuring data integrity and diversity checks. The legal and anonymization model-based conditions for dataset transformations for a given municipality stored at Topicus are defined in 2.1 Anonymization. Next, validity in terms of population representativeness and data diversity is treated in sub-chapter 2.2 External Validity. Finally, data management aiming for the preservation of data integrity is addressed in 2.3 Data preprocessing.

2.1 Anonymization

In every commercial and public domain, aggregated information is collected, stored and processed at data systems. Storing personal records containing confidential data is necessary for administrating processes of citizen or customers. Access and direct insight are granted to a minimum of administrators by means of key encryption. Consequently, data analyses and data publishing violate the privacy of an individual as the likeness of data abuse increases by the number of granted authorizations. This issue is circumvented by any kind of data transformation, including an- or pseudonymization which is subsequently explained. Formally spoken, 'anonymization results from processing personal data in order to irreversibly prevent identification' and should be executed with full conscientiousness along with factors like effort and costs (European advisory body on data, 2014).

2.1.1 Legal Situation

Article 4(1) of the GDPR defines personal data like 'any information which are related to an identified or identifiable natural person' (European Parliament, 2016). Next to the apparent ID's as name, an identification number, location data or an online ID, any data which can be assigned to a natural person. These QI's are for example the telephone, credit card or personnel number. Also, less-explicit information or metadata (see: Chapter 3) falls under this regulation with data about IP-addresses or clocking times.

Recital 26 of the GDPR excludes anonymized data from the principle of data protection. By this, personal information which is not linkable to a natural person fulfills the regulation. Explicitly, pseudonymization and anonymization methods which decrease information content of QI's are allowed. This assertion is extended by including the likelihood of transformed data being re-identified with the currently available technology as absolute privacy do not exist.

2.1.2 Pseudonymization vs Anonymization

Pseudonymization decouples the data subject's context while anonymization holds its context on a global level. By this, an ID is overwritten by a substitute/token which is in most of the cases generated by a hash function which generates randomized strings (e.g. HX57d). This token can subsequently be linked with a decentral database containing (semi)sensitive information. Thus, only the unauthorized access of both databases enables the violation of privacy with, nevertheless, big consequences for individuals.

Anonymization, however, hides the identity of an individual by decreasing QI's granularity without interfering with the integrity of the data subject. Therefore, anonymized datasets can be centrally stored circumventing the obstacle of joining records. Figure 2.1. shows an example of an anonymized data table T^* with the form $(Identifier, QuasiIdentifiers, SensitiveAttribute, NonSensitveAttribute)$.

User ID Name Ethnicity	Date of Birth	Gender	Zip Code	Health Condition	Class
Suppressed	09/**/1964	Female	1902*	Incontinence	C2
	//196*	Male	19022	Incontinence	C1
	//196*	Male	19022	Incontinence	C1
	09/**/1964	Female	1902*	STD	C2
	//196*	Male	19022	Incontinence	C1
	09/**/1964	Female	1902*	STD	C2
Identifiers (Suppressed)	Quasi-identifiers (Generalized)			Sensitive Attribute (Private)	

*Figure 2.1 | Anonymized Table T**

The suppression of the user ID is caused by the property of an ID having assigned an indistinguishable value which cannot be generalized in its granularity. Hash functions can additionally be applied for generating an ID substitute to avoid a drop in the number of bits. This would result in a memory loss as suppression denoted by '*' equals 16 bits and a 5-digit hash as 'HX57d' equals 48 bits. Hence, anonymization for the QI's and pseudonymization for the ID's can ensure a balanced trade-off between data integrity and privacy as ID's can be randomized by hashes and QI's keep their context for readability and data mining purposes.

2.1.3 Threats

Data leakage enables unauthorized persons or organizations to search for patterns in databases for their own purposes (Klösgen, 1995). The analytical process is expressed in the concept of Knowledge Discovery in Databases (KDD). The original variable distributions can be preserved by some anonymization methods for discrete variables where the loss of information can be nearly cut out (Klösgen). By doing this, business analysts can choose based on the attribute properties in a database a suitable anonymization method aiming the optimal trade-off between data utility and privacy protection as illustrated in Figure 2.2.

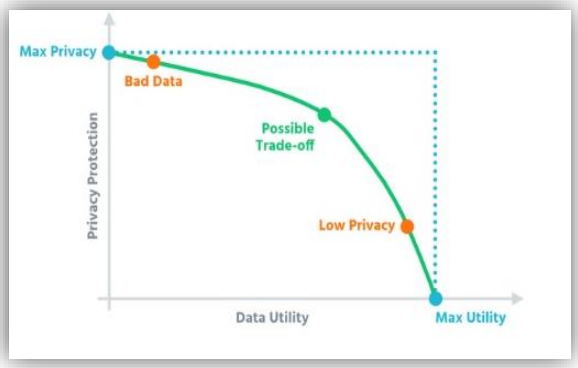


Figure 2.2 | Privacy-Utility Dilemma

Reidentification risks are categorized into two global attacker models, namely linkage attacks and probabilistic attacks (Fung, B.C.M., et al, 2010). The first category assumes that attackers got background knowledge about an individual via another database (e.g. Microcencus or Income Files). Accordingly, sensitive attributes can be derived by linking the additional database to the target database. This direct identification procedure via reference QI's is called Record Linkage. In an appropriate anonymized dataset, individuals cannot be singled out directly and a group of possible candidates with the associated sensitive attributes are present. Assuming a homogeneity of those attributes still allows the attacker to extract precise valuable information. This issue is defined as Attribute Linkage. Next to that, the assurance that an individual with any vulnerable information is part of an anonymized sub-group in target database with the existing background knowledge is expressed in the Table Linkage attacker model. The second category shifts the original sensitive information distribution without significantly changing analysis outcomes by adding noise. Either sensitive information of existing records is modified, or fictional records are added by conditionalized random functions. By this, the probabilistic-based conclusions of an attacker are likely to be incorrect and individuals' privacy can be ensured. Figure 2.3 shows common privacy models with the associated attacker models summarized by the anonymization method review of B.C.M Fung (2010).

Privacy Model	Attack Model			
	Record Linkage	Attribute Linkage	Table Linkage	Probabilistic Attack
k-Anonymity	✓			
MultiR k-Anonymity	✓			
ℓ-Diversity	✓	✓		
Confidence Bounding		✓		
(α, k)-Anonymity	✓	✓		
(X, Y)-Privacy	✓	✓		
(k, e)-Anonymity		✓		
(ϵ, m)-Anonymity		✓		
Personalized Privacy		✓		
t-Closeness		✓		✓
δ-Presence			✓	
(c, t)-Isolation	✓			✓
ϵ-Differential Privacy			✓	✓
(d, γ)-Privacy			✓	✓
Distributional Privacy			✓	✓

Figure 2.3 | Privacy Model Attack Matrix

2.1.4 Non-perturbative methods

The decrease of attribute granularity through generalization is a common procedure for data transformation. Non- perturbative methods are supported by software as ARX Anonymization Tool, Amnesia or UTD Anonymization Toolbox which are used by both science and industry. In principle, one input variable in a privacy model implies no knowledge about the nature of a sensitive attribute which lowers complexity in the conditions. Given the supported privacy models in the software named above, the review is restricted to k-anonymity, l-diversity and t-closeness.

The conditionalized generalization of attributes also known as k-anonymity is the process of aggregating district values to generic values. As each individual is linked to QI's, high-probability association can be drawn resulting in a re-identification of the individual (Byun, J.-W., et al, 2007). This is due the unique combination of the individuals' QI's allowing the coupling with another database via Record Linkage. Therefore, the privacy model k-anonymity was established which set the requirement of the indistinguishability of any record to the remaining $k - 1$ records. In the healthcare domain, a common value threshold value is $k = 5$ which translates to the condition that each sequence of values in table $T(Q_1, \dots, Q_n)$ including n Q's occur at least 5 times in $T(Q)$.

An advanced privacy model named l-diversity was established by A. Machanavajjhalato (2006) to prevent individual identification. Due to homogeneity of sensitive attributes (illness, crimes) associated with generalized QI_n*, the sensitive attribute s of individual n can still be identified. The district l-diversity privacy model requires the occurrence of at least l sensitive attributes in a clustered group of records. The recursive

(c, l)-diversity model extends this condition by requiring that the frequency of the most sensitive value is less than the sum of the frequencies of the $m - l + 1$ least frequent sensitive values multiplying by some publisherspecified constant c, that is, $f_1 < c \sum_{i=l}^{m} f_i$. This constant should be chosen such $c < {}^{1}/_{l}$ to obtain a feasible privacy model.

Lastly, t-closeness tackles the representativeness issue of l-diversity by reflecting the real sensitive attribute distribution as in reality. Suppose one sensitive attribute is dominant in a table T and the remaining sensitive attributes occur only one single time in T. Then an attacker can still single out a record with high probability in a l-diversity privacy model. Therefore, t-closeness defines a frequency range for the sensitive value f_i such it satisfies

$$f_i - t < f_i < f_i + t \quad \forall f_i, 0 \leq f_i + t \leq 1$$

2.1.5 Synthetic data

Randomization is an alternative to generalization. k-anonymity and l-diversity might be disadvantageous for data miners since suppression and generalization lead to a loss in information by means of non-sensitive knowledge and variable distributions (Bayardo, R.J. & Agrawal, R., 2005). While adding noise to an attribute by a uniform randomizer is convenient for the privacy protection, it is less suitable for analyze purposes. Conditional swapping algorithms preserve the associations between QI's and sensitive attributes while maintaining the statistical integrity within an attribute but correlations across attributes are distorted and might be inconsistent. Algorithms for assigning new QI's based on a probability value distribution ensure the same privacy level as in l-diversity or permutation while higher the statistical value of non-sensitive knowledge (Yang, W. & Qiao, S., 2010). This is convenient for numeric values but strings requiring a specific context are hard to generate. For example, overwriting the company 'Bedrijf123 department x' with 'Bedrijf456 department y' is only possible with additional databases which are in most cases either non-existent or not (directly) accessible.

Microaggregation

The perturbative methods described above decoupling most of the entities from its original value. Microaggregation however overwrites only a fraction of the data by the following stages as described by A. Rodríguez-Hoyos (2018):

1. Cluster records by k-anoymity algorithm guarantying that each tuple of key-attribute values is identically shared by at least k records in a data set.
2. Replace all quasi-identifying values by a common representative tuple

By this the inherent information loss is reduced as less frequent and very distinct values do not influence the information content of the remaining records in the cluster. For example, the transformation of the tuples $\{\text{man, unmarried}\}^A, \{\text{man, unmarried}\}^B \& \{\text{other, widow(er)}\}^C$ into the microaggregated form $\{\text{man, unmarried}\}^*$ would result in a lower loss of information than the k-anonymized tuple $\{\text{human, no partner}\}^*$

2.1.6 Systematic Literature Review

Every software development iteration cycle is characterized by a testing phase which demands a certain amount of time. As already described in the literature section, anonymization methods for realistic testing procedures are key to accelerate the software development process. As the optimal trade-offs between data utility and the degree of anonymization is dependent on the attribute characteristics of a database, a systematic literature review might give an inside which (advanced) anonymization methods are suitable for mostly non-binary, character-rich and loosely defined table attributes. Simultaneously to the research question selection, the overall research question is compromised to the following:

> Which anonymization method retains data integrity while providing a low level of individuals re-identification?

An established literature review framework is introduced which provides all necessary steps for contributing to the research question. This guideline is especially applicable in information system (IS) related research topics and fits therefore perfectly the purpose of the given research matter. Webster & Watson (2002) step-by-step approach gives an alternating broad and narrow instruction which is structed like this:

1. Shaping the matter of research by sieving relevant journals' table of contents for keywords & determining boundary conditions for search strings in scientific databases as Scopus and Web of Science. Applied conditions are:
   (a) Search strings: 'Anonymization' AND 'Integrity' AND 'Identification', 'Anonymization' AND 'Utility' AND 'Identification', 'Anonymization' AND 'Consistency AND 'Identification', 'Anonymization' AND ('Consistency OR 'Traceability')
   (b) Keywords: Data privacy, k-anonymity, ℓ-diversity, privacy-preserving, data mining, privacy protection, knowledge preservation, Data Swapping
   (c) Date range: >1995
   (d) Inclusion & exclusion criteria: See Inclusion & exclusion criteria tables
   (e) Title and Abstract: Controlling if the titles and abstract relevant to the research question
2. Scanning content and reference lists of passed articles and include them in the literature list if they contribute to the answer of the research question
3. Developing an overview of existing theories.
4. Preparation an augmented concept-matrix with units of analysis which puts the theories in a two-dimensional matrix and eventually splits them into sub-groups. The matrix should be concept-centric for reader lucidity.

The research question is divided into constructs with refined searching terms which form the foundation of the review. In addition, terms will be extended by synonyms or related terms if applicable.

Constructs	**Related terms**	**Broader terms**	**Narrower terms**
Anonymization	Depersonalization		
Integrity	Utility	Correctness	Consistency
Identification	Traceability		

Table 2.1 | Possible Search Terms

Search strings

After clarification of search terms, the logical connectives are drawn up. As the goal is to find as most relevant articles in different databases, all single terms are searched isolated to find data- or information system-based search results. For example, depersonalization is often used in neuroscience and psychological contexts and are only related to information system in 0.6% of the cases according to the Web of Science database. Next to that, search terms as traceability are compared to 'identification' less used in articles, so the term will occur less in the search string matrix. The first relevance estimation is purely based on the Science

of Web database. The logical connective in the following three-dimensional matrix is '∧' or 'AND' in words except for the consistency and traceability string. Due to lack of search results, the logical connector is '∨' or 'OR' in words.

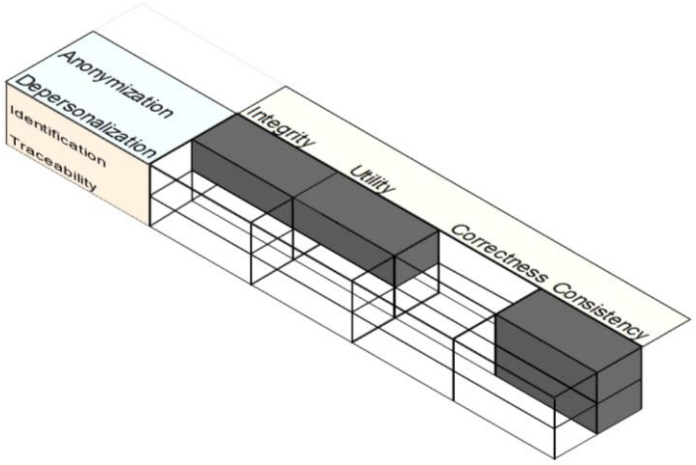


Figure 2.4 | Three-dimensional Search String Matrix

Inclusion & exclusion criteria

In addition to the search strings, some extra key words were added to increase the number of relevant articles. Same applies for exclusion key words with the crucial different to decrease the number of results.

Inclusion criteria	**Reason for inclusion**
Keywords: Data privacy, k-anonymity, ℓ-diversity, privacy-preserving, data mining, privacy protection, knowledge preservation, Data Swapping	These were typical key words in the literature section about anonymization.
Census Microdata	Microdata holds sensitive information about individuals but is most importantly used for linkage to anonymize dataset. Articles which evaluate this possible threat should be included.

Table 2.2 | Inclusion criteria

Exclusion criteria	Reason for exclusion
Subjects: Encryption, Radiotherapy, Oncology, Pervasive computing, Sociology, Arts Humanities	Those subjects roughly focus on efficient anonymization methods.
Key words: alert systems, community detection, GPS (receiver), transactions, neuroimaging, data auditing	Those key words roughly focus on efficient anonymization methods. Implementation differs too much from the research's matter.
Journals: Frontiers in Neuroscience, Social Science & Medicine, ACM Transactions on the Web, IEEE Transactions on Information Forensics and Security	Journals are not rarely privacy focused.
All articles including after-leakage procedures as identity disclosure or data publishing	Not relevant due to perspective change to data thieves.
Pre 1995 articles	First efficient k- anonymity approaches were established in 1998
Non-English, Dutch or German articles	No proficiency to read those articles.

Table 2.3 | Exclusion Criteria

Search results

Three scientific databases were checked to obtain a comprehensive perspective on the body of knowledge associated with the given research matter. The consulted databases are the Web of Science, Scopus & Google Scholar. After compression, it was decided to exclude Google scholar for the literature review since search options are not advanced enough. In order to obtain as most as possible relevant articles, search terms were compromised to the at most the stem if several conjugates are used in the research context. For example, the word 'identification' is reduced to 'identif*'.

Search String	Scope	Date of search	Date range	#Entities
Web of Science				
Anonymization AND Integrit* AND Identif*	Topic or Title	12/1/2019	1995 - present	8
Anonymization AND Utilit* AND Identif*	Topic or Title	12/1/2019	1995 - present	97
Anonymization AND Consistenc* AND Identif*	Topic or Title	12/1/2019	1995 - present	6

Anonymization AND (Consistenc* OR Traceabilit*)	Topic or Title	12/1/2019	1995 - present	15
Scopus				
Anonymization AND Integrit* AND Identif*	Title, Abstract and Keywords	12/2/2019	1995 - present	19
Anonymization AND Utilit* AND Identif*	Title, Abstract and Keywords	12/2/2019	1995 - present	213
Anonymization AND Consistenc* AND Identif*	Title, Abstract and Keywords	12/2/2019	1995 - present	5
Anonymization AND (Consistenc* OR Traceabilit*)	Title, Abstract and Keywords	12/2/2019	1995 - present	31
Gross Total				**394**
Irrelevant articles based on inclusion/exclusion criteria				-356
Removing duplicates				-9
Removed after reading abstract				-22
Included after complete reading				+1
Net Total				**8**

Table 2.4 | Search Results of the Systematic Literature Review

After the final selection, the content of each article was summarized for getting a clear view on the main concepts which were treated in the articles. This table is the basis for the concept matrix.

#Nr.	Author	Title	Content
1	(Sánchez, D., et al, 2014)	Utility-preserving privacy protection of textual healthcare documents	- Two-step **removal** and **generalization** method for semantically related terms which finds optimal threshold between utility (Precision) and privacy (Recall) expressed as F-value based on Web search engine probabilistic functions

			- First step for **generalization** in respect to sensitive terms in the same hierarchy and second step in **across hierarchies** with semantically related terms - **Vulnerability** measured by difference in Human (expert opinion) and method output
2	(Majeed, A., et al, 2017)	Vulnerability- and Diversity-Aware Anonymization of Personally Identifiable Information for Improving User Privacy and Utility of Publishing Data	- Estimation of QIs **vulnerability** by bootstrapping samples from database and building **random trees** (machine learning algorithm) from which data distributions can be divided - Calculation of diversity within sensitive attribute classed by at similar k-dimensions with the **Simpson index** - Algorithm loops until the **least acceptable vulnerability value** for a sensitive attribute
3	(Gal, T.S., et al, 2014)	A data recipient centered de-identification method to retain statistical attributes	- Modified version of **microaggregation** with crucial difference of masking also non-sensitive attributes by generating new value based on covariance models as linear regression - **Synthetic data**
4	(Kohlmayer, F., et al, 2015)	The cost of quality: Implementing generalization and suppression for anonymizing biomedical data with minimal information loss	- Entropy, **Precision and Recall** comparison of syntactic privacy models tested against established clustering algorithms - **Monotonic privacy models** outperform combined models by output quality (utility)
5	(Rodríguez-Hoyos, A., et al, 2018)	Does k -Anonymous Microaggregation Affect Machine-Learned Macrotrends?	- **K-anonymized microaggregation** trough nearest neighbor algorithm maximum distance to average vector algorithm (MDAV) rarely influence the accuracy of Machine-Learned Macrotrends, privacy is not significantly increased

6	(Poulis, G., et al, 2017)	Anonymizing datasets with demographics and diagnosis codes in the presence of utility constraints	- Advanced $(k, k^m) - anonymity$ algorithm applicable when treat knows m sensitive attributes as well as demographics of individual - In addition to normal k-anonymity, the QI of an individual is also non-identifiable from k-1 entities
7	(Fung, B.C.M., et al, 2010)	Privacy-Preserving Data Publishing: A Survey of Recent Developments	- Systematic review between **(non)-permutation anonymization** methods with all common algorithms and based on data metrics as ILost and quantified information/privacy trade-off formulas
8	(Khokhar, R.H., et al, 2014)	Quantifying the costs and benefits of privacy-preserving health data publishing	- Trade-off is chosen on **cost calculations** with risk of privacy breach and the connected monetary value of information loss and costs of potential lawsuits due to attacks

Table 2.5 | Summary of Selected Articles in Systematic Literature Review

Reading the articles give an insight about the main concepts in combination with the research types. The table beneath can be seen as a classification matrix which maps each article in the type of contribution. The category "Algorithm Application" denotes a research with introduced a new (adjusted) algorithm for either a non-perturbative method (generalization and compression) or perturbative method (randomization). 'Sensitivity analysis' indicates a research with measure effects on parameter changes as example indistinguishability range k in terms of utility and privacy formulas within one given algorithm (vertical approach). In contrast, researches which compare serval methods (horizontal approach) in respect to performance measurements as information losses or F-scores are found in the category 'Methods Comparison'.

	Non-perturbative methods		Perturbative methods	Methods Comparison	
#Nr	*Algorithm Application*	*Sensitivity analysis*	*Algorithm Application*	*Quantitative*	*Qualitative*
1					
2					
3					
4					
5					

6					
7					
8					

Table 2.6 | Concept and Research Type Matrix

Afterwards the content and research type tables are unitedly matched into a third concept matrix. As (non)-permutation anonymization methods are sanitizing either the QI's or the sensitive attributes (sometimes both), a category split is made. While the anonymization of QIs danger the consequences of a value, attribute or table linkage attacks, the anonymization of sensitive attributes decreases the integrity of a record. Next, quantification methods are split in F-score [formula which combines Precision (Utility) and Recall (Privacy)], associated costs of a method or other performance measurement methods. Entities marked in grey are less important for the database of Topicus.

		Anonymization model		Measurability approach		
#NR	***Author***	*Sensitive Attribute-focused*	*QI-focused*	*F-value*	*Costs*	*Other*
1	(Sánchez, D., et al, 2014)	Maximum information term from a given tuple extracted from the SNOMED-CT database while fulfilling privacy threshold τ. First iteration with taxonomically related and second with semantically related terms.		Evaluation based on identification rate by method and human experts on anonymized diseases Wikipedia articles.		Utility preservation
2	(Majeed, A., et al, 2017)		Vulnerability for each granularity of			Information Loss determined

			Qis determined by Random forest (RF) algorithm. Formation of generalization levels determined by k-anonymity with checks on diversity and evenness.			by varying values of k. Accuracy of RF algorithm.
3	(Gal, T.S., et al, 2014)	QIs are clustered by basic a basic k-anonymity algorithm and sensitive attributes are masked by linear regression with Qis as independent input variables.		[might be calculated but not done.]		Utility and Privacy measured independently by percentage of sensitive attributes changed significantly .
4	(Kohlmayer, F., et al, 2015)	Three most common used generalization algorithms Flash, OLA, Incognito are compared to five different de-identified databases.	.	F-score (relation between privacy and utility) and attribute entropy is evaluated.		
5	(Rodríguez-Hoyos, A., et al, 2018)	Clustering of QIs determined by nearest neighbor		Sensitivity analysis for		Accuracy measureme

		algorithm deviated from k-anonymity and with average class values (synthetic data)		different k-values.		nt of algorithm.
6	(Poulis, G., et al, 2017)	$(k, k^m) - anonymity$ algorithm where clustering takes place two folded. Generalization with assumption that attacker knows m sensitive attributes and the demographics of individual (both Qis and sensitive attributes are generalized).				Normalized Certainty Penalty (measurement for information loss). Utility Loss (UL).
7	(Fung, B.C.M., et al, 2010)	Mostly qualitative comparison of common (non)-permutation anonymization models regarding privacy models, anonymization operations, information metrics and anonymization algorithms.				Information Loss as theoretical concept - no quantitative comparison
8	(Khokhar, R.H., et al, 2014)	Standard k-anonymity, ∈-differenctial privacy and LKC privacy models algorithm compared from financial perspective.			Cost of anonymization due to information loss. Fines or penalties. Risk of privacy breach.	

Table 2.7 | Detailed Concept Matrix with Quantification Factors

2.2 External Validity

External validity ensures the applicability of findings to other circumstances. In the optimal case, the same result will be obtained using different populations, settings, time points or other operationalization of the same concept. These variables all together lead to the main concept of generalization which is the foundation in every research. Quantifications of results meets with barriers since conclusions over non-researched cases are consistent with an enclosed argumentation and thus, uncertainty about drawing the consequences (Swanborn, 1993). Even if external validity and generalization are seen as interchangeable, a small but crucial difference has to be established for this section. In a nutshell, external validity is be treated as a function of the researcher and the design of the research and generalizability as a function of both the researcher and the user (Ferguson, 2004).

2.2.1 Representativeness

The definition of representativeness is operational. A standalone assessment can never be made since the term should only be used together with data which addresses a given question. Therefore, if the data can answer the question, it is representative (Ramsey, C.A. & Hewitt, A.D., 2005). By taking a sample, a subset of the population can be analyzed where the sampling error is an input parameter for the degree of representability. No error automatically means that the sample is representative.

Population Selection

Before drawing a sample, a suitable population should be chosen. The selection is the matter of research questions and the initial point of statistical analyses. Size and properties of the pre-determined population influence the sampling methods and should be selected to the resources available. For instance, the student population of the whole Netherlands is no suitable reference when conducting a survey about the political attitude in front the Faculty of Arts. A strong bias might be the result.

Sample Size

The sample size can affect the propositional value of a hypothesis. In fact, the probability of the occurrence of extreme values might be higher with a bigger sample which forces the statistician to eventually reject a null hypothesis. Therefore, an appropriate number of samples should be drawn for meaningful conclusions. A guideline for the accuracy is the population size (N) and sample size (n) ratio $\frac{n}{N}$. The larger this ratio, the more accurate the sample will be. Although most studies normalize their sample size, each population should be evaluated along individual characteristics for selecting a suitable number of samples (Coa, Y., et al, 2002). Suitable is in this case an appropriate trade-off between sample diversity and sample size.

2.2.2 Data Diversity

In the domain of software engineering, it is assumed that an application can be implemented in serval ways. This purposed variety is descended from the conjecture that different implementation possibilities are associated with various software designs which should generate different software faults as output (Ammann, P.E. & Knight, J.C. , 1988). This is convenient in testing since user cases can be evaluated and algorithms can be generalized to prevent application errors. Data diversity can be equated with design diversity since it also has a direct influence on the output. Therefore, diverse data is essential for generating adequate failure set.

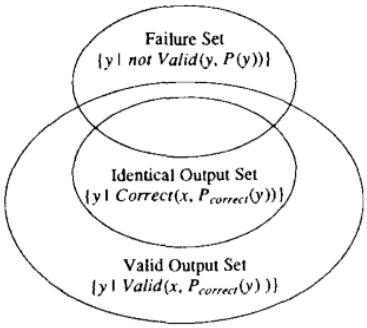


Figure 2.5 | Set in the Output Space for given Input x

Diversity indices

According to the literature, there are several measurements for diversity. Two concepts, both mainly used in ecology, should be evaluated for further specification. Specimen richness s is the number of species present in a population and specimen evenness in the relation between the species in a population. Combined, they form a framework for diversity indices. One of the most common used indices is derived from the Hill family. Here, the true diversity is high if the average proportional abundances in the sample are equal to the proportional abundances in the population (Routledge, 1979). The following formula generates a diversity value N_a in the interval $[0,1]$ where a denote the Hill's order (effective number of equally frequent species in sample), p_i the proportional abundances of the i^{th} type and R the total number of specimen.

$$N_a = \begin{cases} \left(\sum_i^R p_i^a\right)^{\frac{1}{1-a}} & for\ a > 0, a \neq 1 \\ exp\left(-\sum_i^R p_i\, log(p_i)\right) & for\ a = 1 \end{cases}$$

2.3 Data preprocessing

Big Data is accompanied with imperfect data. Data preprocessing is the umbrella term for the transformation of raw into usable data within the domain of the data mining process. This step concerns deleting, adding and, most importantly, changing entities to pre-defined constraints. According to the yearly published Data Science report by Crowd Flower in 2016, the majority of data scientists agree that cleaning and organizing data take the most of their time while being the least enjoyable task of their daily activities. While this phase should not devour a whole research, demonstration of validity and reliability of measurements is crucial (Heerkens, H., et al, 2017). Therefore, automatization is the key word for avoiding unnecessary time wastage. This is achieved by distributed machine learning algorithms which replaced standard single resource algorithms for agilizing the learning process. Using series-connected computers, sophisticated data analyzes still have fast computation times which accelerate the *Knowledge Discovery in Databases* process (García, S., et al, 2014). Switching the perspective from resource-focused approach to a data-focused approach, reveals several possibilities for data preparation and reduction. Possible techniques include data transformation, integration, cleaning or normalization which all aim to lower the complexity of data (Ramírez-Gallego, S., et al, 2017). The table beneath shows serval raw data treatment possibilities.

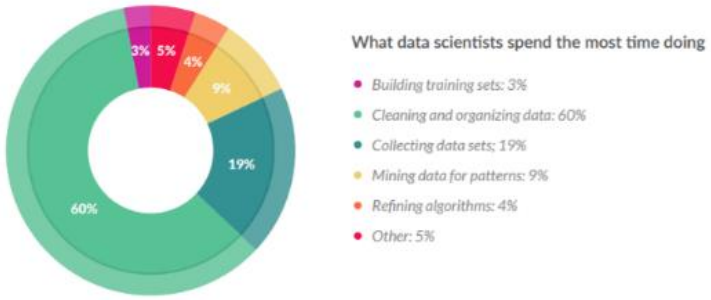


Figure 2.6 | Time Consumption of Data Scientists' Activities

Supercategory	Subcategory	Threats	Solutions
Imperfect data	Missing values imputation	- Poor knowledge - Wrong conclusions - Strong bias	- Discard - Probabilistic models for refill
	Noise treatment	- Input/output influence - Strong bias	- Data polishing - Noise filters
Dimensionality reduction	Feature selection	- High computational costs	- Reducing redundant data
	Space transformations	- High computational costs	- Set transformation into less projections
Instance reduction	Instance selection	- Biased subset selection	- Trial and error
	Instance generation	- Biased subset selection	- Define suitable criteria

Table 2.8 | Preprocessing Data Techniques

Data Quality

While data processing describes the global procedure and possible algorithms for data mining, data quality checks entities along several quality dimensions. The starting point is always a set of raw data which needs to be edited for detecting errors or fluctuations. While some literature provides some broad and general classification types as accuracy, consistency and completeness (Cooper D. & Schindler, P., 2014), most literature categorizes dimensions in a more in-depth way. Here, the central point of view in data quality is intrinsic. However, most organizational problems include a threatening error source namely, data consumers. This group gained more and more choices over their computing environment which asks an extension of the intrinsic view to more categories (Strong, D.M., et al, 2002). High-quality data is meant for the data consumer and can be divided into four categories which are listed in Figure 2.7. Intrinsic data quality is connected to differences in input data which origins from multiple sources. While believability describes the general uncertainty, accuracy is a measurement of the closeness to the real value of a source. Accessibility denotes the degree of technical accessibility by computers and (security) software. Contextual data quality focusses on missing, poorly formatted and loosely categorized data and representational data oriented on the meaning and format of data. There are some dependencies between those dimensions which are evaluated in the following sub-chapter.

DQ Category	DQ Dimensions
Intrinsic DQ	Accuracy, Objectivity, Believability, Reputation
Accessibility DQ	Accessibility, Access security
Contextual DQ	Relevancy, Value-Added, Timeliness, Completeness, Amount of data
Representational DQ	Interpretability, Ease of understanding, Concise representation, Consistent representation

Figure 2.7 | Data Quality Categories

3 Properties Test & Production Database

In this chapter, the composition and properties of both test and production database are evaluated. In 3.1. the Entity Framework of the database is addressed, the nature of data is central to chapter 3.2, the Entity relationship structure is described in 3.3 and representativeness and data diversity are evaluated in chapter 3.4.

3.1 Data Modeling Infrastructure

In software engineering, a fundamental decision is made regarding the order of modeling tasks. A preferred and commonly used framework is the code-first approach when building an application which got also the preference of Topicus. Central to this method is the definition of entities in classes in an object-oriented programming language before migrating them into database tables. The interconnecting data bus between class and table attributes is called Object-Relational Mapping which essentially automates the transfer of object codes to a relational database (Nong Y., 2014). By this, the software engineer has full power over his scripting process as changes in the domain model can be directly implemented. Databases are logically decoupled from the domain model as they function as storage for data only without considering mapping structures generated by an Entity Framework. For structuring purposes, software engineers follow the inheritance concept which groups object classes in two categories, namely (1) superclasses and (2) subclasses. A superclass (e.g. parent) is an overarching class which is directly related to at least one subclass (e.g. children). The Entity Framework converts those classes into database tables and assigns relations to each other which can be either directly done via foreign keys or via coupling pre-defined Associative Entity. The last term is an entity which maps multiple database tables into one reference table through primary or foreign keys. The hierarchical distinction between classes and associative entities improves the logical setup of the scripting process and creates less interconnections between tables but more tables in the relational database. Therefore, clarity and readability of the database has slightly decreased as many tables serve as a reference including ID's only with no actual information content. In respect to the anonymization itself, any sanitization of those ID's should be left out of consideration as SQL reference errors will occur in the implementation.

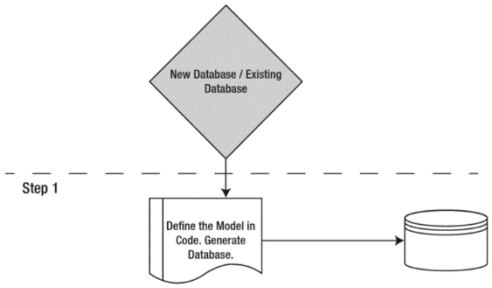


Figure 3.1 | Code-First Approach

3.2 Nature of Data

Running Gidso smoothly requires in its current version a database of in total 605 tables. The application visualizes information about individuals, organizations and related processes as diagnosis or product assignments. Associated information can be grouped in activities, notifications or changes. Quantitatively, inherent information is overly present in the production database and should be treated as insensitive attributes for the anonymization iterations as they cannot identify individuals but rather give indicators for the individuals' global circumstances (e.g. reporting times at *wijkteam* location). This restricts the choice of suitable sanitizable database tables in the following.

Metadata

Data components which refer to superior data are called metadata. The most common form of metadata is descriptive and can be put on a level with supportive or transactional data (Gartner, 2016). A representative example of such entities is shown below.

	id	aanmaakdatum	afzender	berichtidentificatie	ontvanger	client_id
1	646,526,956	2017-06-15 02:11:36	0153	646385853	06290104	646,526,957
2	646,526,967	2017-06-15 02:11:39	0153	646385970	06290104	646,526,968
3	646,526,978	2017-06-15 02:11:42	0153	646386045	06290404	646,526,979
4	646,526,989	2017-06-15 02:11:45	0153	646385686	06290104	646,526,990
5	646,527,015	2017-06-15 02:11:51	0153	646385807	06290404	646,527,016

Figure 3.2 | Database Table 'regio_ibericht_toewijzing'

The table can be identified as an Associative Entity as it has the function to couple ID's while referencing to other tables with (semi)sensitive data. It shows that the municipality has sent a message to an employee of a mental health care organization with the request of assigning related people to the clients' record. Without any context, data from the table can be interpreted such information as individual's location (re. 0153 equals Enschede) or primarily responsible in each mental health care organization (re. frequencies of assignments) can be divided. However, transactional data alone like these can hardly harm the privacy of individuals as the information content is low.

Distribution of data types

Pinning down the information listed in all data tables combined is essential to already pre-clarify anonymization methods with might be applicable. A table containing only Booleans with two states has already theoretically a significantly lower information content than a table which only accept integers. Gidso has many fields where loose descriptions can be added containing multiple sentences allowing numbers and (special) letters. For every character n entered in a text field, the theoretical combinations for distinguishability are 68^n. This means that a random generated string with 8 characters can be differentiated from $68^8 - 1 = 4.57\ x\ 10^{14}$ possible generated strings. Consequently, column attributes with a high degree of uniqueness or distinguishability from which information networks can be divided are highly vulnerable to data leakages. An information network is the conjunction of data which unitedly can be deployed for dividing information in a given context. For example, leaked decentral Customer ID's are without any associated personal information little useful for hackers in comparison to central insurance policy numbers which might be coupled to a second leaked database of a different company. A rough overview of the data type distribution is shown below which exhibits a string-rich database.

Data type	Frequencies
String	High-Medium
Numeric	High
Boolean	Low-Medium
Time	Low-Medium

Table 3.1 | Data Type Distribution of region-independent Database

3.3 Entity Relationship Structure

Content-wise, the 605 tables are separated into four main functional table categories. Stating with the two smallest categories, namely 'fact' and 'agenda' (In English: 'appointment'), the binary status of the patient record (active or non-active) and appointment details are archived subsequently. Both categories contain mostly attributes with foreign keys to other tables where personal information is stored. The category 'form' stores choice list arrays which are visualized in Gidso. These data records are not related to personal information as entities are unassigned. The most essential and base category is called 'regio' (In English: 'region') and stores personal information as well as sensitive attributes. In the anonymization, these data tables get the majority of attention as represents about 80% of the whole database environment.

Naturally, there are foundations on which detailed data tables are based. The data table with most connections is 'regio_organisatiemedewerker' but acts more like an associative entity with no actual readable information listed. In contrast, the data table 'regio_doel' contains at least the objective of the clients' treatment procedure and is, therefore, suitable for anonymization. However, this table must be merged into a table containing personal information before any procedure can take place. As the figure already shows, the ER-Model visualizer uses the IDEF1X notation for the dependencies to other data tables. For merging tables consisting of organized .csv files, one-to-zero-or-more relationships need special attention as several entities per client must be compromised in an array before connecting the associated information to a reference table. For example, a client entity with several disease ID's listed in multiply rows must first be merged into a single-cell array before connecting this information to a hospital location. Otherwise, a loss of information in the transformation is the result.

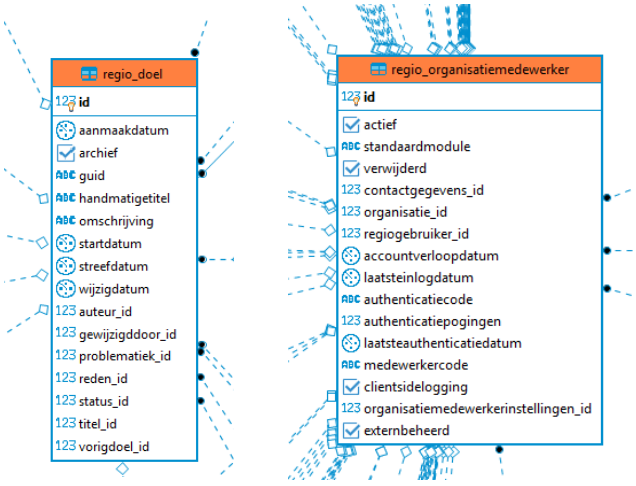


Figure 3.3 | Key Connections to remaining Tables

3.4 Representativeness & Diversity

At this moment, two kinds of relevant databases coexist at the servers of Topicus, namely the testing and the production database. As derived in the first chapter, the current testing database is assumed to be inconvenient in terms of representativeness and diversity. By adding context to the usage of Gidso, it immediately gets clear that empty tables for making appointments, single entities for locations and nearly no activity changes are not consistent with the real-world situation. For retaining integrity, this significantly smaller testing database should have the same size as the original database because entities with certain QIattributes cannot be plotted in histograms or cumulative distribution functions. This due to the unique character of addresses, names or BSN's. Therefore, the smaller testing database is per definition not as diverse and likely not as representative as the production database. Nonetheless, measurements are taken to undermine this assertion.

Methodologically, the assessing of the data-related properties should take place at different levels of granularity. In a global picture, the row count estimator in the database administration tool is a reliable indicator for the percentage of filled tables as single-sided precluding of observations do not follow the principles of sampling. The table to the right shows there is a clear lapse of data which is caused by absent input in Gidso. An example for such a single-sided lack of data can be found in the table 'regio_bestand' where authorizations, triages or just notifications are stored as .pdf files with foreign keys to the client's personal information. While in the production database, 161,411 records are listed, the testing database has zero entities and is therefore empty. To open up the dimensions of entity disparity, selected tables are the basis for measurements of the desired indicators.

Database	**Percentage Fill**
Production	$^{378}/_{605} \approx 62.5\%$
Test	$^{240}/_{596} \approx 40.3\%$

Table 3.2 | Table Fill Percentage

Representativeness

A clasping issue occurs when choosing representative tables reflecting the representativeness of a database. Requirements for a founded table selection are set up and chosen above a random selection procedure. This is due to the enormous number of tables which function as Associative Entities and have no strings or contextual numeric data. For the comparison, a reference population with associated QI's is necessary which is accessible from a census. Sensitive and specific information as psychological issues or unhealthy lifestyle can be hardly estimated as either no official documentation is available or unreported cases cause a bias in the comparison. Therefore, only personal information tables apply to a certain comparison outcome. The most detailed tables with personal information for the stakeholders 'Client' and 'Relations' are listed in the overview below. Data is acquired from the Centraal Bureau voor de Statistiek *(In English: Central Agency for Statistics)* as a reference population for Dutch citizen. Attributes for the analysis are male-female ratio, percentage of feasible BSN's *(Elfproef),* average age (or parental authority age) and the civil status ratio (married/unmarried).

	Male-female ratio		**Average age**		**Civil Status ratio**		**Elfproef**	
Test Database	Test	CBS	Test	CBS	Test	CBS	Test	CBS
Personengegevens	23.57%	49.65%	32.63	42.00	∞	0.555	100%	100%
Relatiegegevens	-	49.65%	-	58.65	-	0.555	-	100%

Table 3.3 | Test Database Representativeness Comparison

	Male-female ratio		Average age		Civil Status ratio		Elfproef	
Production Database	Prod.	CBS	Prod.	CBS	Prod.	CBS	Prod.	CBS
Personengegevens	46.68%	49.65%	44.69	42.00	0.355	0.555	99.99%	100%
Relatiegegevens	49.08%	49.65%	63.8	58.65	0.265	0.555	100%	100%

Table 3.4 | Production Database Representativeness Comparison

The comparison in the table above illustrates that the testing database is filled with semi-logical entities as BSN's were generated by an algorithm instead of a randomizer. However, attributes as the civil status are sparsely filled and only 5 of the 250 entities had a state assigned which is always labeled as 'unmarried'. Therefore, the married/unmarried ratio is ∞. The table 'regio_relatiegegevens' does not even contain one entity in the testing environment and is consequently kept out of consideration.

In contrast, the production database represents the Dutch Citizen more accurately even if the reference population is adjusted with subject to the age in the 'regio_relatiegegevens' table as relations are likely to be parental authorities with the average age of 58.65 years. This increased number is estimated by the average age of citizen aged 25 or older. Since this table is not coupled to any record date, the day of birth is only an indicator for the age as non-active client records from a few years ago slightly increase the average age.

Diversity

Measurements in diversity differ in the property that they do not need any context as observations are only compared within a given population (in this case: within a data table). Thus, treatments or locations entities can, accompanied by personal information, be used for the diversity comparison. As has already been mentioned, records are absent in most data tables and therefore, only tables are chosen with sufficient entities in both databases. As a consequence, tables are picked out of three sub-domains, namely:

1. Appointment
2. Nature of problem
3. Personal information (QI's)

The Hill numbers provide quantitative insights into the uniqueness of a dataset. An increasing number order q increases the weight of dominant entities in the calculation of the diversity index as can be divided from the following step-for-step definition. While the Hill numbers of order 0 are the total numbers of attribute entities, the Hill numbers of order 1 are roughly the number of frequent occurring entities. For order 2 the Hills Numbers consist of less frequent occurring entities (Nagarajan, 2018).

Analogous to the representative measurements, suitable attributes should not be randomly chosen but rather selected from different domains. The associated attributes from the pre-defined sub-domains listed above are 'location of appointment', 'problem domain', 'family name' and 'civil status'. Divided from the literature section, the diversity index is applied to civil status attribute from the test database for demonstration purposes.

$$Hill\ Numbers_{Test,civil\ status} = \begin{cases} R = 6 & for\ a = 0 \\ exp\left(-\sum_{i}^{R} p_i\ log(p_i)\right) \approx 2.38 & for\ a = 1 \\ \left(\sum_{i}^{R} p_i^a\right)^{\frac{1}{1-a}} = 2 & for\ a = 2 \end{cases}$$

By computing the remaining tuples $\{database\ environment, attribute\}$, Figure 3.4 shows the graphs with their typical convergence curves. A steep curve indicates an entity diversity while a flat curve indicates entity homogeneity. The Hill Numbers are shown on both y axes with normalized scales to compare the slope of both test and production database.

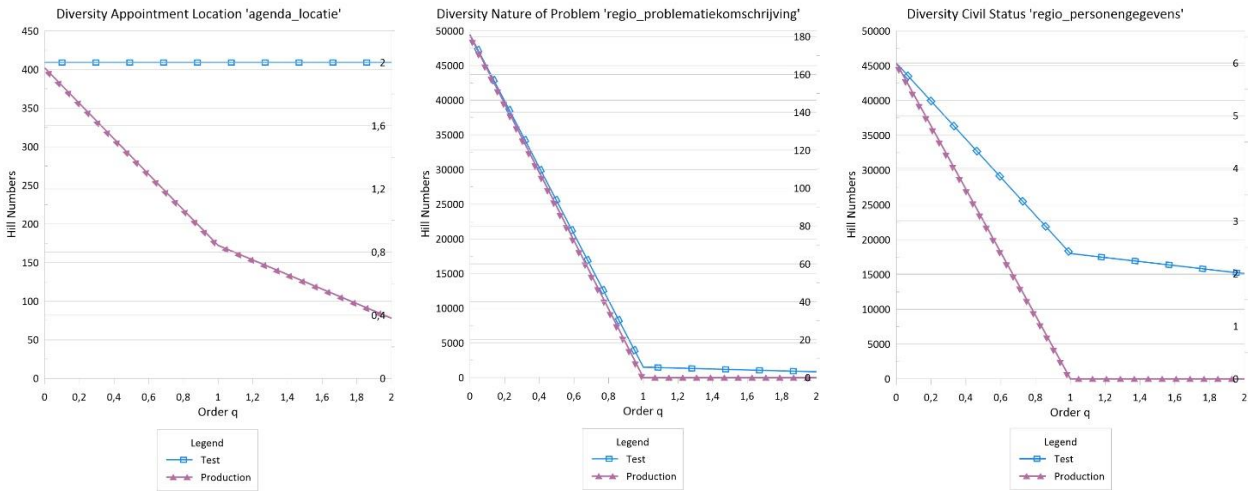


Figure 3.4 | Diversity profiles plotting Hill numbers ${}^{q}D(\infty)$ as a function of order q

All attributes in the production database environment comprise a higher degree of diversity where mainly the first drop from R to Hills numbers of order 1 is remarkable. Especially the curve of the attribute 'Appointment Location' shows this discrepancy exemplarily as the only two values entered in the whole testing database are 'Enschede' implying pure homogeneity. In the context of search queries in databases, a higher degree of diversity automatically results in a longer response time. Having relatively few unique values relives index-based searches and speed up the response time. Therefore, data diversity is added as a measurement variable in the anonymization iteration.

4 Design & Development

This chapter specifies the usage and processing of data along with measurement variables for the anonymization. Chapter 4.1 addresses the Dataset selection, in chapter 4.2, data filtering and quality checks are executed, chapter 4.3 specifies generalization hierarchies, chapter 4.4 approach suitable privacy models in ARX and chapter 4.5 determines sanitization parameters for those models.

4.1 Dataset selection

In order to satisfy the legal obligations of the GDPR, ID's referencing to data subjects as names or identification numbers are the object of QI-focused anonymization methods. As definitions for personal data are kept global and should be interpreted as broadly as possible, an automated search query only with the commonly known ID's is avoided. The logical consequence is a systematic but manual perusing of the filled 378 tables. In the same turn, (semi)sensitive data and text fields are identified and placed into an overview.

Furthermore, tables containing personal data and tables containing sensitive data are coupled with each other by establishing table relationships to prevent information loss. Consequently, clients related to multiple treatments require multiple columns and are accordingly processed. The example beneath shows a typical join procedure.

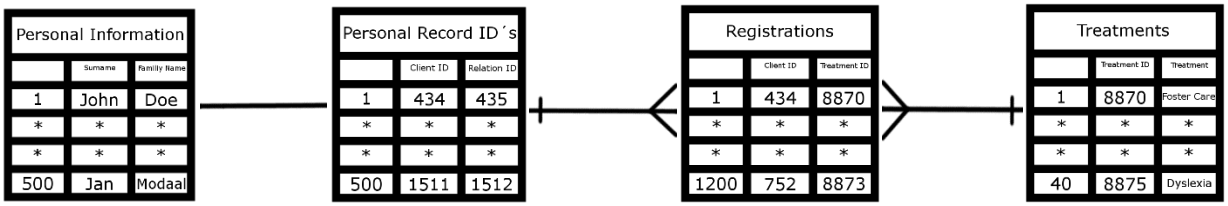


Figure 4.1 | Table Entity Relationships in Joining Context

Figure 4.1 highlights the property that every registration is accompanied by a single treatment entity with no restriction to the Client ID frequency. However, the number of sensitive attributes for a given client is set to a maximum of 5 to prevent high computation times and too many columns in one table.

Sensitive terms detection

Every table is systematically checked for ID's like Name, BSN, Address, Birthdate, E-mail, Phone, Status, Lifestyle & Education by running a global search query. Next to that, (semi)sensitive attributes and rare ID's are detected manually as only a few repetitive attribute names exists. Exemplarily, standardized attribute names as 'beschrijving' *(In English: description)*, 'toelichting' *(In English: explanation)* or 'probleem' *(In English: problem)* must be mentioned. Also, attributes are pre-categorized in semi-sensitive and sensitive attributes. This is done by my own and the company's estimation as the threshold of sensitivity is always subject to the interpreter and obviously to the law. These estimations serve the purpose of selecting suitable tables for the sanitization only and not as definite ready-to-use anonymization scheme as metadata (which falls under personal information according to the GDPR) is not always clear to appoint. The sanction for an all-compounding implementation should be separately done by a data right lawyer.

Regarding the selection, an overview matrix is set up with a separation for the three stakeholders, namely: Client, Relations & Third Parties. This classification serves to identify tables with the maximum number of QI's or Sensitive Attributes which play a central role in the whole database and are most complete. Table 4.1 shows the simplified exemplarity structure of such a matrix for the category 'Client'.

	ID		*QI's*		*Sensitive*		*Semi-Sensitive*	
Table	**Name**	**BSN**	**Address**	**Birth date**	**Description**	**Support**	**Description**	**Support**
1		X	X				X	
*	*	*	*	*	*	*	*	*
50	X	X			X	X		

Table 4.1 | Simplified Sensitive Term Search Matrix

One key result of this identification process is the lack of couplable sensitive attributes to the personal data of relations. The suitability for the following anonymization iteration is therefore obsolete. However, if personal data remains unchanged, the re-identification risk rises also for linked clients as relations may live in the same household or have the same family name. After all, the sum of QI's is the highest in the following merged tables.

1. Regio_personengegevens via Regio_client via Regio_voorziening_beschikking via Regio_voorziening_voorziening to Regio_product
2. Regio_fo_aanmelding via Regio_fo_aanmelding_betreft to Regio_problematiekonderwerp
3. Regio_betrokkene_standaard

These combinations are renamed and put into the summarizing table 4.2.

Dataset	Number of Joins	Entities (raw)	QI's (number of distinct values)	Sensitive attribute
PRODUCT_PERSOON [PRPE]	4	46217	sex (2), civil status (6), birth date (1877), date of death (45), birthplace (981)	Product (57)
AANMELDING_PROBLEEM [AANP]	2	71867	sex (3), birth date (13887), residence (13), ZIP-code (31558), street (14533), mobile number (48099)	Problem (17)
BETROKKENE [BETR]	0	68962	e-mail (1550), organization (176), mobile number (1453)	Relation Description (331)

Table 4.2 | Selected Data Tables for Anonymization

4.2 Data Filtering & Data Quality

As all kinds of users have access to Gidso, people have all kind of interpretations and preferences about data entry. This makes grouping or clustering of attributes challenging because of either small variation in writing style or the different interpretations of the level of detail. Also, absent data is a basic problem which normally results in dropping those records. Therefore, data manipulation is expedient to improve the quality of anonymization.

Data filtering is the most straightforward tool for data manipulation. The underlying assumptions are the necessity of a sensitive attribute for every record. This removes between roughly 21 and 97 percent of the records. This seems logical when looking at the choice and context of the selected data sets. Being listed in the personal information table does not necessarily mean that a specific product is required if it, after the first meeting, transpires that self-support is more useful. Next to that, the 97 percent decrease in the third-party dataset is also reasonable due to the lack of requirement to fill in the function of the notifier.

Achieving a high degree of contextual data quality is done by checking the completeness of all selected QI's and systematic spelling checks. For example, format changes in phone numbers are applied to strengthen consistency (053 – 544 & 053 544 are transformed to 053544) and writing styles of companies are put on the same level (companies with apartment name or 'BV' added or removed). The preservation of integrity in the context of intrinsic data quality gives the last polish to the dataset. This can be achieved by coupling existing databases from the CBS to a few attributes. By this spelling mistakes of city names or municipalities and most

importantly integrity checks on the feasibility of ZIP-codes in the Netherlands are revealed. This last step leads to the following dataset sizes.

Dataset	**Original**	**Sensitive Attr.**	**Contextual Data Quality Check**	**Intrinsic Data Quality Check**
PRPE	46217	23647 (-48.8%)	21577 (-8.8%)	21577 (±0%)
AANP	71867	56334 (-21.4%)	563 (-99%)	530 (-5.9%)
BETR	68962	1965 (-97.2%)	1879 (-4.4%)	1879 (±0%)

Table 4.3 | Dataset Quality Checks

4.3 Clustering Hierarchies

Hierarchies are essential for clustering purposes. This context is added to an entity to minimize the number of distinct values at a rising hierarchy level. In other words, the granularity is decreased by a high hierarchy level. A birthplace can, for example, be split into a range of precise to global levels as shown in Figure 4.2.

Buitenzorg	{Noord-Holl...	{West-Nederl...	{Nederland}	{Europese Un...	{Wereld}
Ten Boer	{Groningen}	{Noord-Ned...	{Nederland}	{Europese Un...	{Wereld}
Tavas	{Tavas}	{West-Turkije}	{Turkije}	{Azië}	{Wereld}
Ordu	{Ordu}	{Noord-Turki...	{Turkije}	{Azië}	{Wereld}
Malang	{Malang}	{Oost-Java}	{Indonesië}	{Azië}	{Wereld}
Rome	{Rome}	{Midden-Itali...	{Italië}	{Europese Un...	{Wereld}
Bolsward	{Friesland}	{Noord-Ned...	{Nederland}	{Europese Un...	{Wereld}
Zeven	{Niedersachs...	{Noord-Duits...	{Duitsland}	{Europese Un...	{Wereld}

Figure 4.2 | 6-Level Hierarchy of Birthplaces

Most hierarchies are essentially divided from existing databases from the CBS. Unfortunately, only the current 380 municipality hierarchies in the Netherlands are available which means that historic municipalities, villages and places outside the Netherlands must be added manually. This applies to 605 of the total number of 3630 distinct birthplaces in the dataset PRPE. The same procedure is set up for the ZIP-codes which was more straightforward as all client addresses for registering at a healthcare location in Enschede are also in the region of Enschede. However, the computation times for allocating all house numbers h & ZIP-codes z into a specific neighborhood and area are high with in total $\sum_h^H \sum_z^Z 1 = 7{,}716{,}327$ entities.

4.4 Assignment of Privacy Models

The range of privacy models is subject to the choice of the anonymization software. In an early stage, preference was granted to the anonymization tool ARX as it is established in the industry, open source and it covers most of the non-perturbative privacy models. Synthetic methods are also considered for ID's as BSN's or Names using a conditionalized allocation from reference databases containing standard first and surnames. However, the resulting record-based data utility is per definition low and conflicts with the overall research question. Consequently, the research is mainly based on the output of the software and exceeded by synthetic methods for certain attributes.

The anonymization tool ARX offers a platform where data is transformed along conditions expressed in privacy models. Analysis on utility and privacy is divided in related main variables including (1) *Precision* [Generalization intensity], (2) *Record-Level Error* & (3) *Recall* [Average prosecutor re-identification risk]. These variables will be further explained in the chapter.

A sanitization can either be QI-based only or be accompanied by conditions on the sensitive attribute. Therefore, it is chosen to include k-anonymity and l-diversity respectively as they reflect the just referend options. T-closeness is in theory also a possible candidate but anonymizations iterations are not possible with $\geq 50\%$ degree of suppression due to the l-diversity advanced rule of reflecting the original sensitive attribute distribution in the clustered group. Besides that, an uncommon method is chosen which optimizes the transformation along with an average re-identification risk without considering the (extreme) outliers. It follows that a small fraction of records might have a unique tuple of QI's. This might be advantageous if the risk distribution has a positive skew with a spiky peak and a long tail where only a low number of the records are critical values. This model is called average re-identification risk.

Within every privacy model, every QI can be clustered following the generalization method described in section 4.3. or clustered firstly and be overwritten with the dominant value in the subgroup following the principle of microaggregation. The generalization intensity a 6-level hierarchy for a given clustered QI as in Figure 4.2. of {Enschede, Zwolle, Enschede} into {Overijssel}* for the value 'Enschede' would be:

$$\sum_{z \,\epsilon\, \{Enschede, Zwolle, Enschede\}} Gen.\ Intensity(z) = \frac{(|G| - 1) - g_z}{|G| - 1} \approx 0.8$$

where g_z is the hierarchy level of the tuple z and $|G|$ is the total number of hierarchies for a given QI. The record-based error is zero in this example. On the other side, the same cluster in the microaggregation method would overwrite 'Zwolle' into the dominant value {Enschede} with a generalization intensity of 1 (no suppression) but with a record-based error of 0.33. Therefore, both clustering methods are included in the anonymization iterations.

4.5 Sanitization Parameters

After providing the first baseline, QI's are weighted according to the likeliness of data suppression. Distinct values as the date of death or mobile numbers which probably can identify a client fast are weighted lower in the model. This is due to the restrictions of both k-anonymity and l-diversity as keeping those distinct values results in a higher degree of generalization in the remaining QI's. On the other side, parameters with a detailed hierarchy as for the birthplaces are weighted higher to prevent dominant suppression.

The exemplary parameters for the dataset 'PRPE' are shown in Figure 4.3.

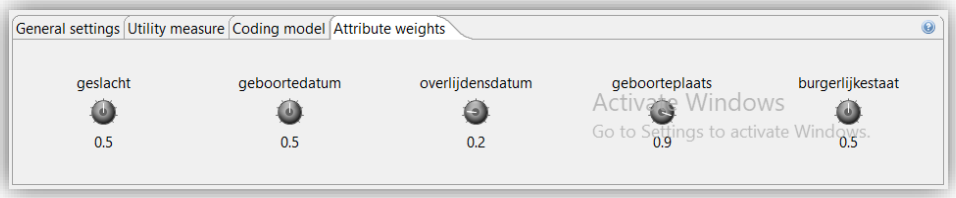


Figure 4.3 | Attribute Weights for the Dataset PRPE

Besides that, the sanitization parameters for the privacy models are based around industry standards to show variable sensitivity. For k-anonymity the parameters are $k = 3$ and $k = 7$, for distinct l-diversity the parameters are $l = 2$ and $l = 4$ and for the average re-identification risk the thresholds are 2% and 5%.

4.6 Outcome variables

In the previous sub-sections, the terms utility and privacy were mentioned. The concept of generalization intensity as a measure of *Precision* was clarified but other outcome variables were left in the dark. Utility in this research revolves also around data diversity expressed in the *Average Class Size* and data integrity expressed in *Record-Based Errors*. For discernment, these variables are further broken down. The *Average Class Size* addresses

the average number of records in a clustered group. While big clusters with high homogeneity reflect relatively low data diversity, small classes do the opposite. *Record-Based Error* is the rate of records for which the context is decoupled by simply overwriting or suppressing data.

On the side of privacy, the *Recall* [Average prosecutor re-identification risk] is the most relevant measurement. Distributions of minimal and high risks are provided by ARX with associated extreme values. For the anonymization iterations, the average risk provided for each privacy model with the extension of providing the upper standard deviation 2δ for the average re-identification model as no risk threshold is predefined. For illustration, a risk distribution for records with (maximal) risk is provided in Figure 4.4 with the maximum re-identification risk of 50% satisfying the condition of a 2-anonymity privacy model.

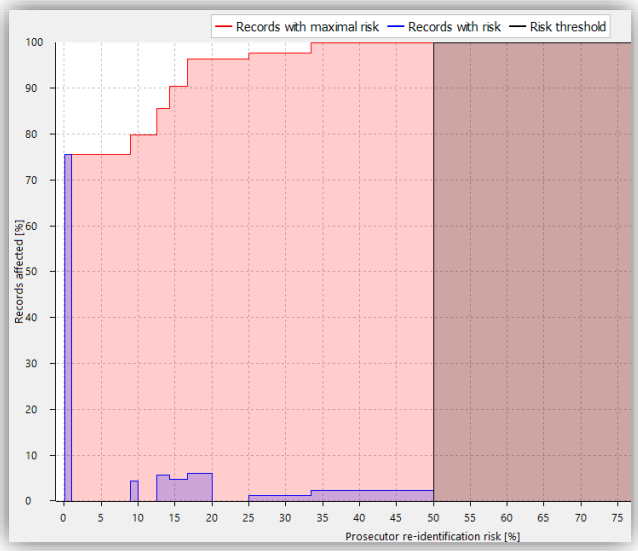


Figure 4.4 | Risk distribution of a 2-anonymity Privacy Model

The trade-off between privacy and utility as addressed in chapter 2.1.3 is quantified by the so-called *F-score*. It provides a harmonic mean of Precision and Recall and contributes to the choice of the most suitable privacy model. The corresponding formula is shown below.

$$F - score = \frac{2 \times Recall \times Precision}{Recall + Precision}$$

A low *F-score* indicates a balanced relation between Recall and Precision while a high score represents the opposite. However, the final choice should not be completely based on these metrics as variable sensitivity might differ crucially.

5 Evaluation

In this chapter, the evaluation of the anonymization iterations is conducted along with the allocation of pseudonyms in chapter 5.1 and pre-defined measurements for the selected privacy models in chapter 5.2.

5.1 Allocation of Pseudonyms

ARX is not capable of generating hashes, let alone readable first and surnames. Existing open access databases containing popular (Dutch) names allows the replacement of existing names via common spreadsheet functions. By applying conditions as same sex and same initials, the real name distribution is represented even more precisely. For the PRPE database, the original initials and the sex were taken for the reallocation. Figure 5.1. shows a selection of popular first names.

Initial	Sex	Name
A	MAN	Adriaan
A	MAN	Adrianus
B	MAN	Bart
B	MAN	Benjamin
D	MAN	Daan
D	MAN	Dave
T	VROUW	Tessa
V	VROUW	Valerie
V	VROUW	Vera
W	VROUW	Wendy
W	VROUW	Willemijn
Y	VROUW	Youssra

Figure 5.1 | External Dataset of Popular Dutch Names

The newly generated names are as readable as the original names and contain roughly the same number of total characters. The composition of two or more first names might not be ordinary but fulfils its function as a pseudonym and can be seen in Figure 5.2. While names are the most obvious ID's, other numeric ID's as social security numbers, location data or online ID's can be conditionally generated without any reference database.

Sex	Initials	1st	2nd	3rd	4th	1st Pseudonym	2nd Pseudonym	3rd Pseudonym	4th Pseudonym
MAN	ZAJ	Z	A	J		Zakaria	Adrianus	Jens	
MAN	LMHS	L	M	H	S	Laurens	Marco	Hidde	Sam
VROUW	A	A				Adriana			
MAN	IHA	I	H	A		Ivo	Hidde	Adrianus	
MAN	FA	F	A			Frank	Adrianus		
VROUW	SAH	S	A	H		Sarah	Adriana	Hilde	
MAN	RBI	R	B	I		Robert	Bart	Ivo	
MAN	JM	J	M			Jacob	Michael		

Figure 5.2 | Newly allocated names based on sex and initials

5.2 Privacy Model Comparison

Names and according pseudonyms make up only a small fraction of a detailed record. Hence, the evaluation is mainly directed to the measurements which were generated by ARX. These are regarded as guidelines for choosing an appropriate privacy model for the aspired level of privacy and the distribution of risk. The following measurements are chosen such a balanced F-score (possible trade-off between privacy and utility) can be derived and data diversity, implying realistic queries, can be rated.

5.2.1 Utility measurements

The previously described utility variables are bought into a relation with the degree of suppression defining the threshold of dropping less-frequently occurring values. By an increasing suppression rate, common values are not forced to be generalized. The Precision & Error data points are given for the suppression rates {5,10,25} on the left and right side of Figure 5.3. respectively. The abbreviations in the legend should be read as {MeasurementVariable_ParameterPricacyModel_TranformationMethod}.

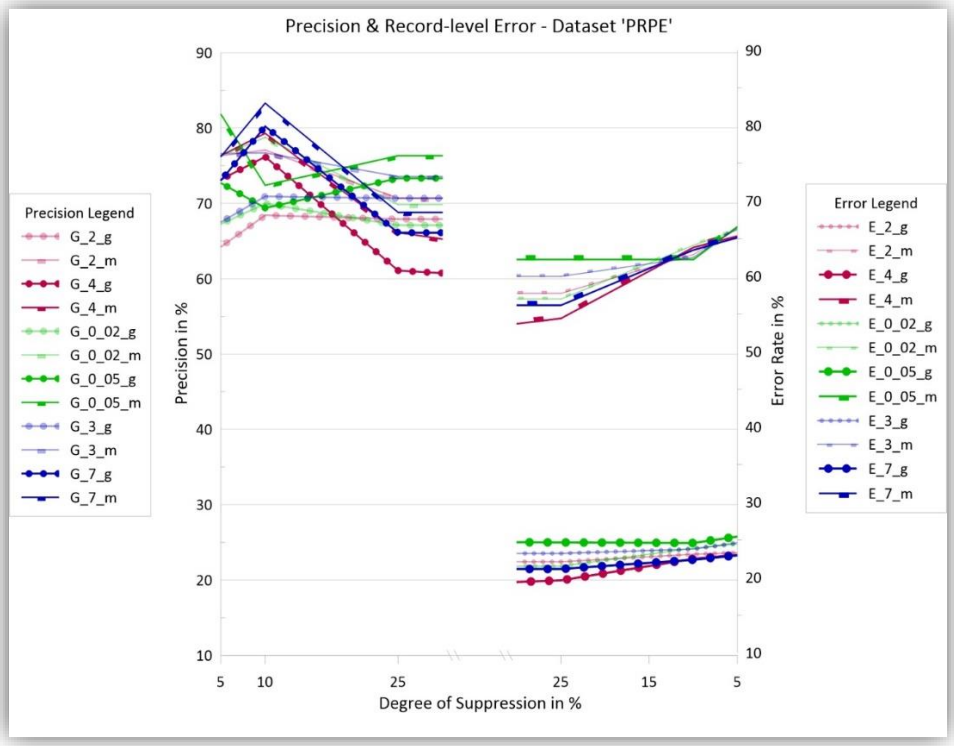


Figure 5.3 | Precision & Record-Level Error related to the Suppression Rate

In terms of Precision, a humped curve with its maximum at 10% suppression is present for the major privacy models and parameters. Two outliers show an opposite behavior which are ascribed to the average re-identification risk with 5%. This happens due to the unrestricted risk distribution where a distinct record with a re-identification risk of 100% is not forced to be suppressed if the average re-identification risk of the dataset is $< 5\%$. In each case, microaggregation transformations have a higher Precision level than generalization transformations within the same privacy model and parameters. In dataset 'PRPE', the generalization intensity is increasing by a stricter privacy model regarding k-anonymity and l-diversity which contrasts the curve shapes of datasets 'AANP' and 'BETR'. Due to relative record homogeneity of 'PRPE', class sizes are relatively bigger, and the clustering algorithm does not interfere with any indistinguishability limit of {2,3,4,7}. The highest Precision is achieved by **mircoaggregated 7-anonymity** with a value of 83,305%.

The record-level error is generally higher for microaggregation as it allows the data transformation to overwrite data. Generalization transformations are less error-prone with **generalized 4-diversity** as the lowest record-level error (18,878% at 25% suppression).

5.2.2 Privacy measurements

Privacy is expressed as the average re-identification risk. Existing privacy model abbreviations and suppression rates remain unchanged from chapter 5.2.1. In addition, the upper 2δ limit covering 95% of the records risks is given for the datapoints {5,10,25} in Figure 5.4 for the average re-identification model.

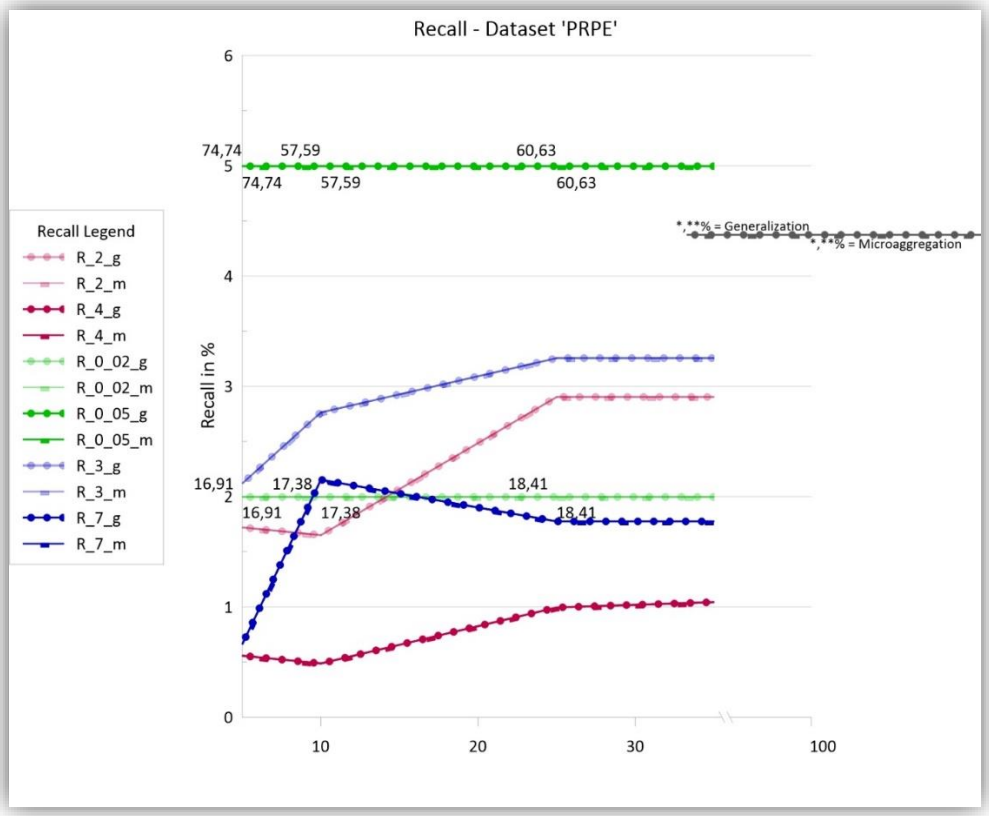


Figure 5.4 | Recall related to the Suppression Rate

An overall consistency is found regarding the Recall curves in all three datasets. Figure 5.4 shows a constant line for the average re-identification model as this is the global restriction. The risk spread is immense with 5% average re-identification and peaks with a value of 74.74% for 2δ. For **7-anonymity**, 5% suppression and for **4-diversity** 10% suppression ensure the lowest average re-identification risk for both microaggregation and generalization. This behavior is consistent for all datasets 'PRPE' 'AANP' and 'BETR'. The upper re-identification risk limit for those two privacy models is $^{1}/_{MP} \times 100\ \%$ where MP $\epsilon\ \{2,3,4,7\}$ is the model parameter.

5.2.3 Class Diversity

In this chapter, the average class size of the untransformed datasets is bought into relation with the average class size of the transformed datasets. A fraction of 100% indicates no change in data diversity. Privacy model abbreviations and suppression can be obtained from chapter 5.2.1.

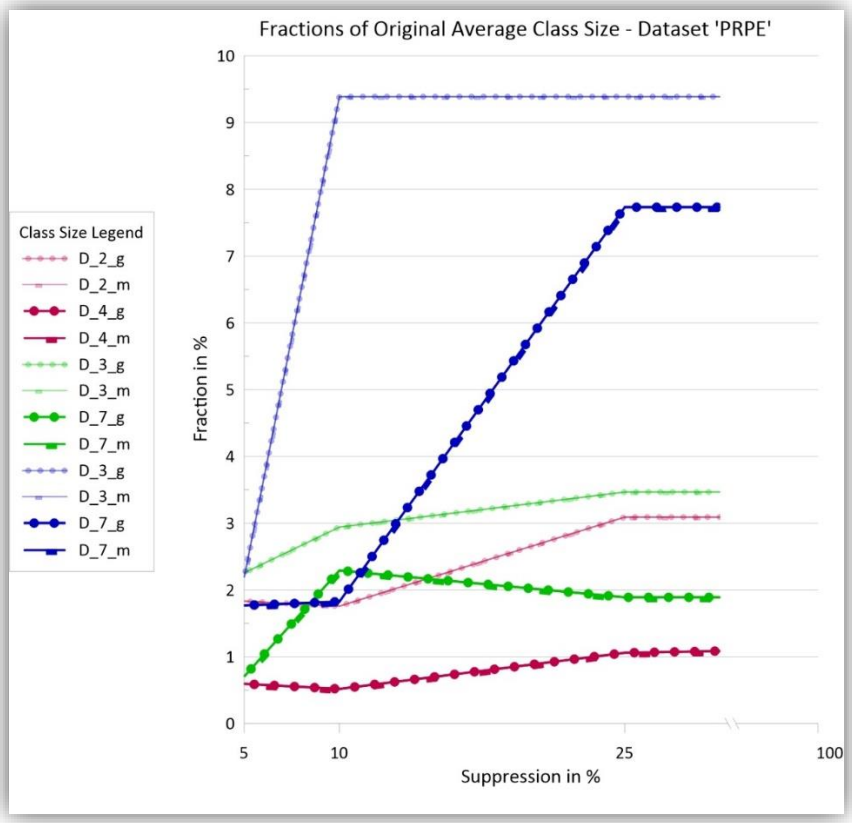


Figure 5.5 | Fraction of Average Class Size related to the Suppression Rate

Naturally, privacy models with fewer restrictions score better in class diversity. For **generalized 3-anonymity**, an original class size of 1.06 is transformed into a class size 10.73. In dataset 'AANP', generalized and **microaggregated 2-diversity** has with 22.82% of the original class size at 25% suppression relatively the highest data diversity. It is noticeable that higher levels of indistinguishability in the privacy models tend to have its highest value at a suppression rate of 25%. A dominant Hill number of order $q = 2$ might explain this shift to the right side of the graph as, for example, the suppression of 25% of very distinct QI's, do not force frequently occurring QI's to be generalized. By this, an absolute maximum of 75% of the original class size could be obtained for one QI and one sensitive attribute.

5.2.4 F-score

The F-score sketch out the harmonic balance between data utility and privacy. Again, privacy model abbreviations and suppression remain unchanged and can be obtained from chapter 5.2.1.

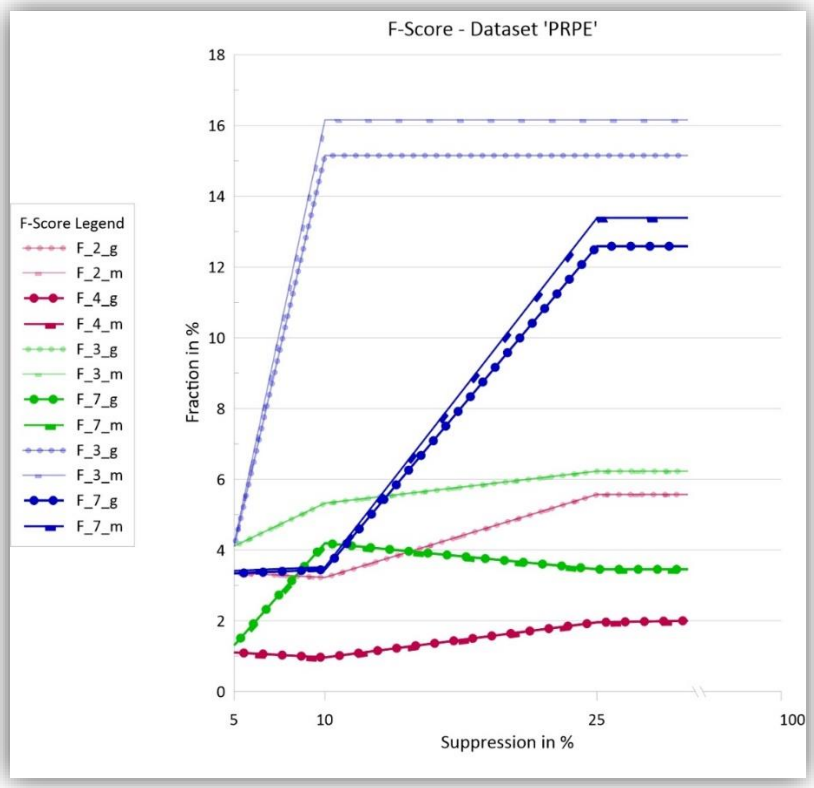


Figure 5.6 | F-score related to the Suppression Rate

By an increasing rate of suppression, the F-scores get large which implies a relative inharmonic mean between Precision and Recall. While 4-diversity holds its balance until a minimum of 10% suppression, the rest has its lowest value at 5% suppression. The same curve behavior can be seen in the remaining datasets where diverse data with a lot of distinct values shows a slight shift to right regarding the extreme values. Also, privacy models with low sanitization parameters (e.g. 2&3) are much more sensitive to changes in suppression compared to higher parameter values.

An absolute and consistent minimum for a certain privacy model cannot be derived as data properties in the given datasets dissimilar. In 'PRPE' and 'AANP', **4-diversity** generates the lowest F-score with 0.97 (10% suppression) and 3.34 (5% suppression) respectively. For the dataset 'BETR', the optimal value can be derived at 1.91 at 25% suppression for **7-anonymity**

6 Conclusion & Recommendations

In this chapter, the quantitative results of the anonymization are qualified along with different scenarios. In chapter 6.1, the conclusion is drawn regarding bug fixing and attribute property-based anonymization, chapter 6.2 addresses recommendations for the anonymization itself but also for data management and in chapter 6.3, an insight for further research is given.

6.1 Conclusion

Ensuring conscientious data handling is anno 2020 only feasible by standardized guidelines. Especially, IT-companies have a special responsibility as they are constantly surrounded by sensitive data. Any sanitization of (frequently) accessed data warrants the protection of privacy and is, in principle, desirable. This research specifies the variable sensitivity of privacy models for the data environment of Topicus. For the sake of clarity, the answer of the research question provides the theorical and practical steps for implementing privacy models to the whole testing environment which consequently solves the action problem. In this section, the research question is listed and answered by considering different source scenarios.

> "Which privacy model provides a harmonic mean of data integrity and re-identification threat aiming the reduction of bug frequencies in software testing by the preservation of realistic and diverse data?"

Naturally, there is no universally valid answer for one specific privacy model. This is due to the link of input and the behavior of the two measurement variables data integrity and re-identification threat. Therefore, the answer is distinguished by two source scenarios addressed in section 6.1.2.

6.1.1 Bug fixing

As derived from the research objective defined in the DSRM, the current testing environment enables the frequent occurrence of falsified query times, display errors or wrong indexing. The comparison of testing and production database showed a disparity in size, representativeness and diversity solving implicitly the cause of bugs. The proposed solutions do not differ in size but in representativeness expressed data integrity and in diversity expressed in the fraction of the original class size. For the three datasets which are used in this research, the desired data properties are met to a degree which is determined by the re-identification treat level. Consequently, an implementation implies a reduction of bug occurrence frequencies caused by the error sources named above.

The expert opinion in section Appendix A. underlines this inference as the executed data handling in combination with evaluated privacy models enables a full implementation with the current workforce of Topicus. Still, the consideration of acquiring external experts for the execution is dependent on the willingness of other divisions to apply anonymization methods to their databases. By quantifying the daily work of a Software Engineer, a percentage of 10 to 25% is directed to maintenance including the work of bug fixing. Therefore, it is a reasonable prospect that the proposed privacy models accompanied by a consequent full implementation improve the time spent on bug fixing.

6.1.2 Attribute Homogeneity vs Heterogeneity

Before implementing a privacy model, a data set should be globally distinguished in its attributes. The Hill's Numbers from section 3.4 guide the categorization.

Homogeneity

If values belonging to a certain QI are largely nearly identical, the column values are said to be (relative) homogenous. This can be quantified by the application of the Hills Numbers as done in chapter 3.4. For datasets, where the value D from q=0 to q=1 and q=2 it remains rather constant than strongly declining, the following conclusion is derived.

For the most homogenic dataset 'PRPE', both 5% average re-identification risk at 5% suppression and 4-diversity at 10% suppression score best regarding the F-score. The scores for microaggregation or generalization do mostly not differ in most of the measurements, but in 5.2.1 a minor disparity emerges. When data integrity is the focus, the favor should be given to the generalization. In contrast, for better generalization intensities, the favor should be given to microaggregation. Regarding privacy, the average re-identification risk model has in comparison to the other privacy models, a big risk spread. For 'PRPE', <2.5% of the records can be re-identified with a risk >75%. Still, the majority of records maintains high integrity and low re-identification risks. Naturally, the data diversity rises with a declining indistinguishability number, which qualifies 3- or even 2-diversity as a suitable candidate for the harmonic mean.

The slightly less homogenous dataset 'AANP' also generates a good F-score for 7-anonymity. However, a comparison of the Recall graph and F-score graph shows a strong correlation and is not considered as fully harmonic. Therefore, the previous conclusions remain unchanged.

Heterogeneity

In contrast, a majority of distinct values for a given QI are considered as a diverse attribute. A strong decline of the Hills numbers from q=0 to q=1 and 1=2 quantifies this diversity.

Dataset 'BETR' is filled with company names and job functions. Thus, the records are little standardized which compliances the (automated) set-up of a generalization hierarchy. In such a case, high indistinguishability restrictions make the data nearly useless and uniform. For example, generalized 7-anonymization has a generalization intensity of 16.54% at 10% suppression. That means that on a 5-level hierarchy, every record is overwritten to the most general hierarchy level and some values are dropped in additionally. Therefore, a favor is given to microaggregation as data diversity is generally higher. By comparing Precision and Recall, the 5% average re-identification risk or 2- or 3-diversity (5% suppression) for microaggregation provide a mean with a slight focus on data diversity. From the privacy perspective, microaggregated 7-anonymity and 4-diversity score better on the Recall measurement but lack of data integrity.

6.2 Recommendations

6.2.1 Privacy Models

The conclusion already points out suitable privacy models. However, in the context of Topicus, there are several influencing variables which supports the decision regarding the appropriate privacy model selection. Naturally, data integrity is important for every data miner but as the current and future testing database will mainly be used by software developers, data diversity is an equally important decision variable. Hence, a diverse search query is likely to be called from the disk instead of the memory and takes more response time. This reflects reality more precisely and is therefore, more time representative. Therefore, **microaggregated 2- or 4-diversity** is suitable in most of the times depending on the data input homogeneity.

6.2.2 Hierarchy generation per data type

Anonymization software can automatically generate numeric hierarchies for every interval. For dates, the set-up is straightforward and are little time-consuming. However, most data are stored or function as a string within a given context. Birthplaces, job functions or ZIP-codes need (external) hierarchies for keeping a part of their information content. Some hierarchies can be accessed via the CBS as, for example, the ZIP-codes and Dutch municipalities, and coupled to the present dataset. For complex and distinct attributes, as birth places outside the EU or Dutch job functions, data input should be standardized, hierarchies should be

generated by Topicus or external hierarchies should be acquired. A possible 3-level job function hierarchy might look like this:

Level-0	Level-1	Level-2
Sociotherapeut	Therapeut	Behandelaar

Table 6.1 | Exemplary 3-level Job Function Hierarchy

6.2.3 Implementation execution

This research addresses three explanatory datasets with varying QI's but naturally, the database holds much more sensitive information. Therefore, a search matrix is provided and added in Appendix 7.2. It is recommended executing anonymizations for at least the listed data tables. Join operations as done in chapter 4.1 are not always necessary as ID's for sensitive attributes are processed in the same way as the readable string. For example, 'Regio_fo_aanmelding' should only be joined with 'Regio_fo_aanmelding_betreft' as l-diversity always needs the coupled sensitive attribute for its model restrictions. Anonymization purely on tables where personal data is listed is not enough in this case. Next to that, metadata can be vulnerable if background information is known by an unauthorized person. Therefore, from case to case, any metadata should be evaluated against possible re-identification risks.

6.2.4 Data input restrictions in Gidso

During the data table selection phase, sensitive information was detected at neutral input fields. Users namely tend to give summary client descriptions containing Names, BSN's, telephone numbers etc. at text fields which are meant for problem descriptions. Although tables containing those data were identified in the search matrix with the highest conscientiousness, there may be undetected sensitive information in unlisted tables. Still, this concerns unstructured re-identifications risks as the majority of users use the input fields correctly. For the future, those text fields should be provided with a warning of **not** entering unnecessary personal information as the entity is already linked with a personal record.

Besides, a drop-down box for relations to the client (e.g. work, friend, mother) or job descriptions as explained in 6.2.2 could help standardizing data and, thus, lower the loss of information during anonymization.

6.2.5 Internal collaborations

The importance of data protection is, partly due to this research, likely to be awakened for Topicus. Still, only a part of these results may be communicated to other divisions due to the company structure. As Topicus deals in every domain with personal data, a common data anonymization approach should be applied in every division. Therefore, software and methods should be identical for guaranteeing consistent data protection across the whole Topicus organization.

6.3 Discussion

6.3.1 Relevance

The starting point of the evaluation is the literature review in Appendix 7.1 which reflect the scientific reference of my research question. While the papers of e.g. G.Poulis (2017) and F.Kohlmayer (2015) are technically well-executed, they address the properties of the input data restrainedly. The introduction of attribute diversity measurements is helpful for the selection of suitable privacy models and related parameters. From an argumentative point of view, this conclusion is accompanied by the fact that homogenous entities overlap which results in bigger initial class sizes. Therefore, higher indistinguishability parameters can be chosen without lowering the data integrity. The results of the three datasets indicate this behavior but cannot demonstrate it significantly. Therefore, this relation should be investigated in future research.

Next to that, the uncommon average re-identification model showed parameter sensitivity for a given dataset. By this, re-identification risk distribution can be initially analyzed and adapted for other privacy models. For example, an 5% average re-identification model with a high-risk spread implies low data integrity for high restricted privacy models as 7-anonymity. This due to the large cut of the 'risk tail' whose records are either highly generalized or suppressed. This allows a bottom-up choice of the sanitization parameter instead of top-down selection which gets the preference is most of the researches.

6.3.2 Limitations

ARX and its capabilities

The anonymization software offers a solid platform for non-perturbative anonymization methods. All common privacy models are available but more recently developed models or algorithms are not selectable. In particular, more advanced methods also considering existing background knowledge of the attacker which makes relatively protected data vulnerable. For example, (k, k^m)-anonymity provides a privacy algorithm

when treat knows m sensitive attributes as well as demographics of an individual. Next to that, data transformation along missing perturbative methods as randomization complicate method comparison as measurement variables are not overlapping in each anonymization software.

F-score as weighting factor

As indicated in chapter 6.1, the Recall variable is very dominant in some datasets. As most of the re-identification risks are in the range (0,5] and Precisions are in the range [17,84], a value increase in of the Recall measurements has a bigger impact on the harmonic mean that the value increase of the Precision measurement. Therefore, alternative weighting formulas might be more meaningful.

6.3.3 Further research

The GDPR also lists pseudoanonymization as a suitable data transformation method. Purely for the execution of search queries, any type of diverse data can result in a realistic response time. Also, a high level of privacy is guaranteed due to hashes as placeholders. Anonymization, however, deals consistently with a suppression rate which results in a full information loss. Therefore, both methods have the potential to become complementary for the case that a harmonic mean between utility and privacy occurs at a high suppression rate. Thus, the combined usage should be further investigated.

Next to that, the privacy models in ARX do not consider background information as a potential threat. As described in chapter 6.3.2., there are existing privacy models for those cases but the link to input diversity is not formed yet. By the adoption of methods in chapter 3&4, the impact on privacy in comparison to the common privacy models can be estimated and judged along with extra effort and costs.

Bibliography

Ammann, P.E. & Knight, J.C. . (1988). Data Diversity: An Approach to Software Fault. *IEEE Transactions on Computers 37(4)*, 418-425.

Bayardo, R.J. & Agrawal, R. (2005). Data privacy through optimal k-anonymization. *21st International Conference on Data Engineering.* Tokoyo: Institute of Electrical and Electronics Engineers.

Byun, J.-W., et al. (2007). Efficient k-Anonymization Using Clustering. *DASFAA* (pp. 188-200). Heidelberg: Springer-Verlag.

Coa, Y., et al. (2002). Comparison of Ecological Communities: The Problem of Sample Representativeness. *Ecological Monographs 72(1)*, 41-56.

Cooper D. & Schindler, P. (2014). *Business Research Methods.* New York: McGraw-Hill/Irwin.

European advisory body on data. (2014, April 10). 0829/14/EN. *Opinion 05/2014 on Anonymisation Techniques*. Brussels.

European Parliament. (2016). *Protection of natural persons with regard to the processing of personal data and on the freemovement of such data, and repealing Directive 95/46/EC .* Brussels.

Ferguson, L. (2004). External Validity, Generalizability, and Knowledge Utilization. *Journal of Nursing Scholarship (Volume 36)*, 16-22.

Fung, B.C.M., et al. (2010). Privacy-Preserving Data Publishing: A Survey of Recent Developments. *ACM Computing Surveys 42(4)*, 1-53.

Gal, T.S., et al. (2014). A data recipient centered de-identification method to retain statistical attributes. *Journal of Biomedical Informatics*, 32-45.

García, S., et al. (2014). *Data Preprocessing in Data Mining.* Heidelberg: Springer.

Gartner, R. (2016). *Metadata: Shaping Knowledge from Antiquity to the Semantic Web.* London: Springer.

Heerkens, H., et al. (2017). *Solving Managerial Problems Systematically (First edition ed.).* Groningen: Noordhoff Uitgevers BV.

Khokhar, R.H., et al. (2014). Quantifying the costs and benefits of privacy-preserving health data publishing. *Journal of Biomedical Informatics 50*, 107-121.

Klösgen, W. (1995). Anonymization techniques for Knowledge Discovery in Databases . *Proceedings of the First International Conference on Knowledge Discovery and Data Mining* (pp. 186-191). AAAI Press: California .

Kohlmayer, F., et al. (2015). The cost of quality: Implementing generalization and suppression for anonymizing biomedical data with minimal information loss. *Journal of Biomedical Informatics 58*, 37-48.

Machanavajjhala, A., et al. (2006). L-diversity: privacy beyond k-anonymity. *22nd International Conference on Data Engineering (ICDE'06).* Atlanta: Institute of Electrical and Electronics Engineers.

Majeed, A., et al. (2017). Vulnerability- and Diversity-Aware Anonymization of Personally Identifiable Information for Improving User Privacy and Utility of Publishing Data. *Sensors* , 1-23.

Nagarajan, M. (2018). *Metagenomics.* Kasaragod: Academic Press.

Nong Y., T. W. (2014). *Developing Windows-Based and Web-Enabled Information Systems.* Boca Raton: CRC Press.

Peffers, K., et al. (2007). A Design Science Research Methodology for Information Systems Research. *Journal of Management Information Systems*, 45-77.

Poulis, G., et al. (2017). Anonymizing datasets with demographics and diagnosis codes in the presence of utility constraints. *Journal of Biomedical Informatics 65*, 76-96.

Ramírez-Gallego, S., et al. (2017). *A survey on Data Preprocessing for Data Stream Mining: Current status and future directions.* Amsterdam: Neurocomputing 239 .

Ramsey, C.A. & Hewitt, A.D. (2005). A Methodology for Assessing Sample Representativeness. *Environmental Forensics (Volume 6)*, 71-75.

Rodríguez-Hoyos, A., et al. (2018). Does k -Anonymous Microaggregation Affect Machine-Learned Macrotrends? *IEEE Access 6*, 28258 - 28277.

Rodriguez-Repiso, L., et al. (2007). Modelling IT projects success: Emerging methodologies reviewed. *Technovation* , 582–594.

Routledge, R. (1979). Diversity Indices: Which Ones are Admissible? . *Journal of Theoretical Biology 76(4)*, 503-515.

Sánchez, D., et al. (2014). Utility-preserving privacy protection of textual healthcare document. *Journal of Biomedical Informatics 52*, 189-198.

Satariano, A. (2018, May 18). Retrieved from New York Times: https://www.nytimes.com/2018/05/06/technology/gdpr-european-privacy-law.html

Strong, D.M., et al. (2002). Data Quality in Context. *Communications of the ACM 40(5)*, 103-110.

Swanborn, P. (1993). External validity abandoned? . *Quality & Quantity*, 211-215.

Topicus. (2020, January 10). Retrieved from Topicus: https://topicus.com/about-us

Yang, W. & Qiao, S. (2010). A novel anonymization algorithm: Privacy protection and knowledge preservation. *Expert Systems with Applications (Volume 37)*, 756-766.

Appendix

A. Expert opinion – Yoshi Koen

Anonymization

Question: Absolute privacy should always be the strive but is however never completely achievable. The GDPR names (pseudo)anonymization as appropriate data transformation method among factors as cost and effort. How would you qualitatively estimate the willingness of Topicus to invest in an all-compounding (pseudo)anonymization for frequently used data in favor of software convenience?

Answer: I would say in so much that willingness towards a certain goal can be qualified, the willingness of Topicus to both invest in software quality and in the proper protection of data is laid down in our recent efforts (and success) to conform to ISO standards 9001 and 27001. Conformity to ISO 9001 involves having software quality as measurable goal, which in turn means investing in methods of ensuring software quality. Among these methods is the exploration of anonymization of customer data in an attempt to solve the natural contradiction between ensuring software remains performant in real world scenarios and one of the ground rules of software data protection: to keep testing environments and production data separate. In this, our willingness to invest in conformity to ISO 9001 and 27001 shows qualified willingness to invest in researching solutions that support the goals lined out in these standards and, if they show both promising and financially viable, implementing them.

Question: Do you consider various anonymization methods for different divisions (e.g. Business Development & Software Engineering) as reasonable?

Answer: I definitely believe a tool should fit the job. As different divisions within an operation (company) seek different sub-goals, I find it only reasonable that different methods would suit different divisions. As we have seen during the research, there are various ways and degrees to approach anonymization with according trade-offs. I would not assume one way of anonymization to fit all use cases, nor would I presume that when anonymization of customer data proves to be promising and viable for one division, it also naturally means that it is a promising and viable solution for all divisions.

Implementation

Question: What are your thoughts on the ease of implementation by the provided data tables, the step-by-step instruction, and the resulting measurements?

Answer: Ease of a task shall always depend on the complexity of the subject matter, quality of the proposed solution and relevant skill and knowledge of those who are to implement the solution. In this case the subject matter is highly complex - as demonstrated by the fact that to date, no single be-all-end-all solution has been embraced in the software development domain. The proposed solution is well-documented, and the relevant skill and knowledge level – especially regarding the mathematical aspect of the solution - is low. One takeaway for me is that while given ample time and effort we would be able to fully implement the proposed solution with the current workforce of Topicus Social Services, it would likely be more financially viable to employ someone with a better suited skillset to fully implement the solution in our operation and/or share resources with other divisions of Topicus as many of our subdivisions have similar software stacks and, undoubtedly, similar challenges.

Question: The unnecessary bug fixing process caused by the current testing database can be eliminated or fastened with more realistic databases. Can you qualitatively or quantitatively estimate the approximate time savings per month for a Software Tester due to fewer bug fixing?

Answer: Noting that it is quite difficult to accurately predict time not spent as a result of a implementing solution rather than measuring time spent for not implementing it, I would approximate that anywhere between 10% to 25% of our time spent on maintenance (i.e. all development work that is not directly related to the new features) might in some way be related to testing against unrealistic data. Do also note this involves not only the time spent software testers but also by software developers, service desk, consultancy and other parties involved with software problems that require handling.

Software adaption:

Question: In the healthcare domain, the background and circumstances of a client are necessary to provide a personalized treatment. Do you think that standardized text fields for problem descriptions reduce the degree of personality drastically?

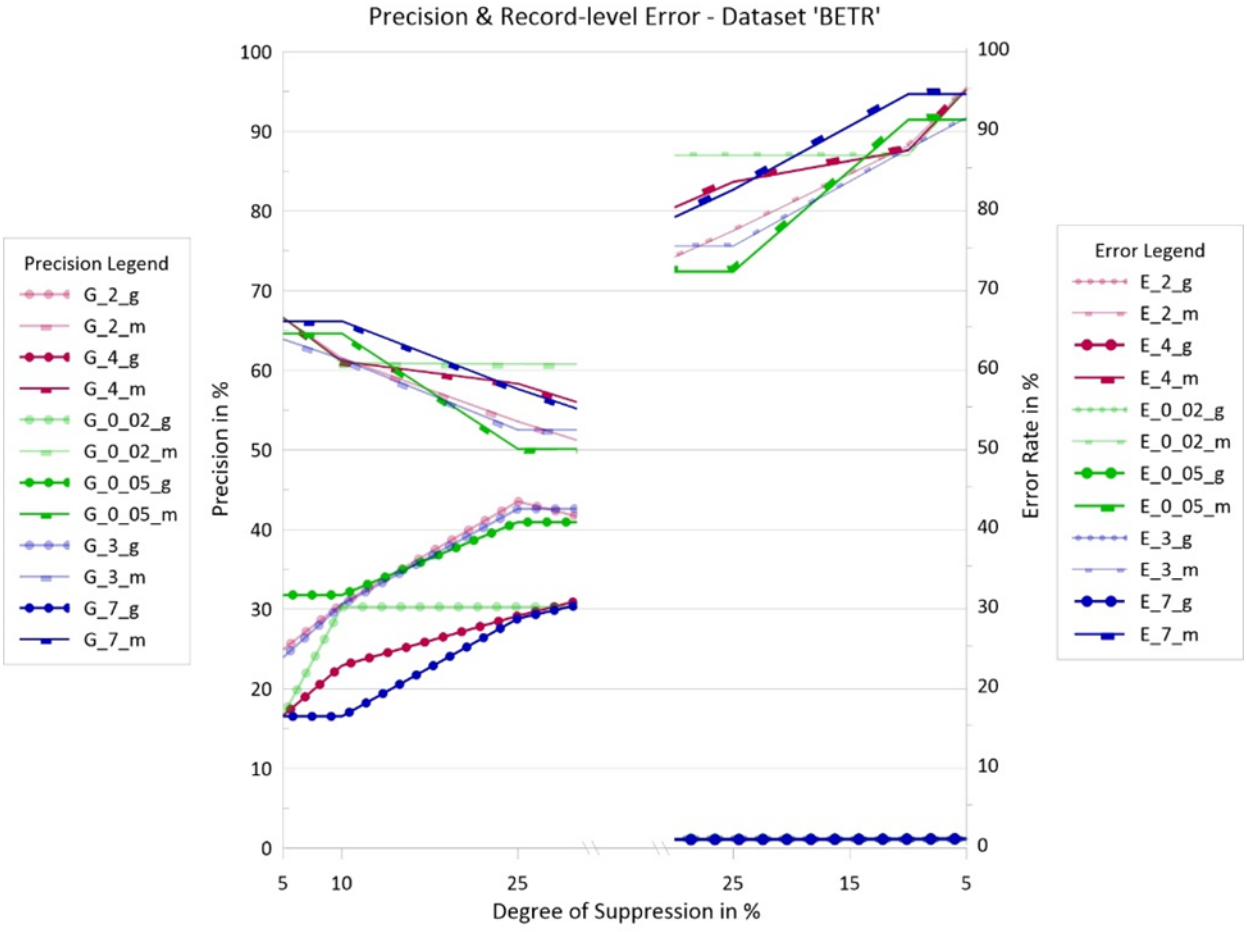
Answer: I would definitely agree that limiting an answer set reduces the degree of personality for almost any given situation. I must note however that how much the degree is reduces strongly depends on context and the data in question, and that reducing personality has no intrinsic value until the sum of all reduction is

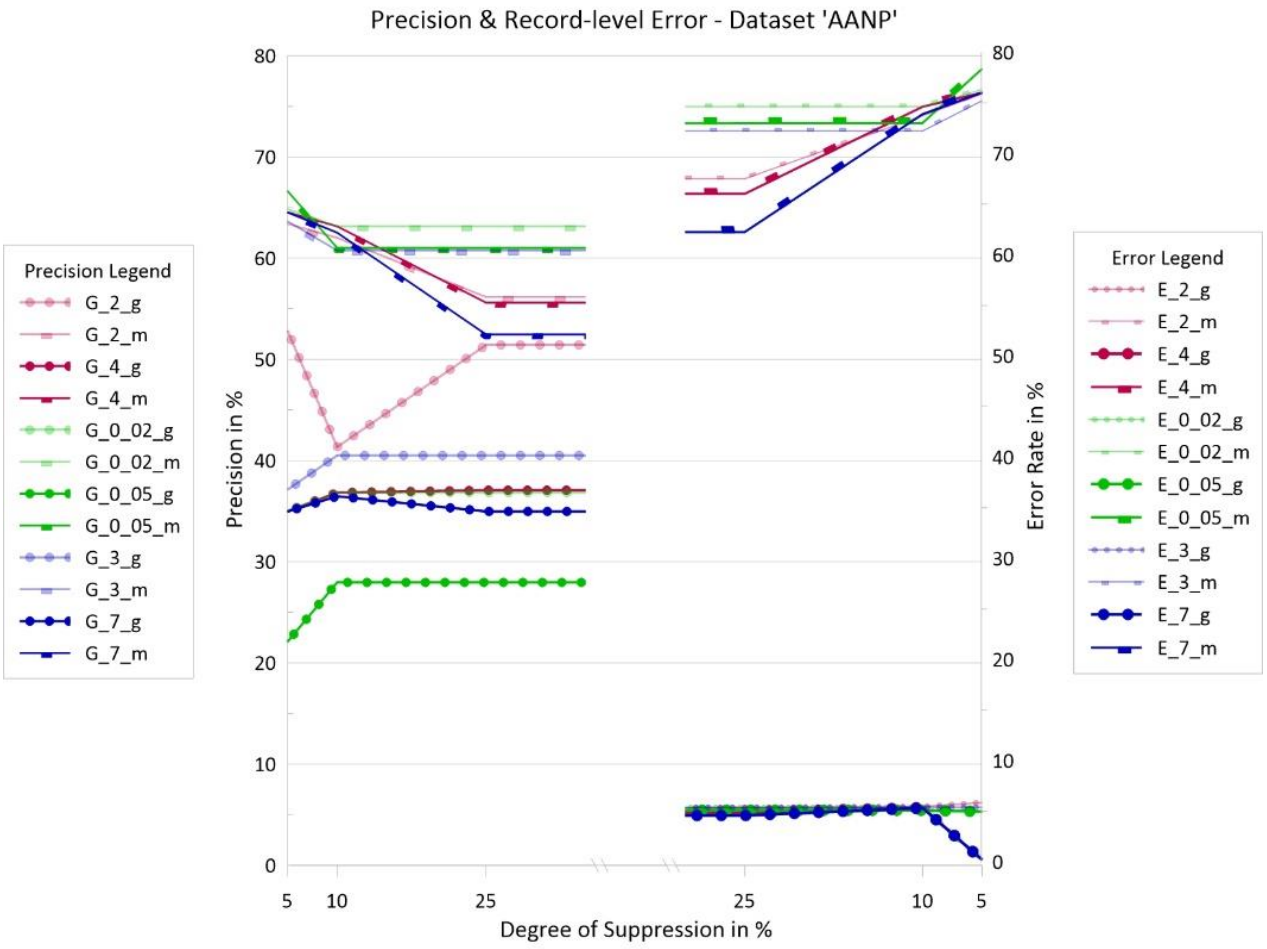
sufficient to make the entire set of information non-reducible to a single person by parties who should not be able to do so.

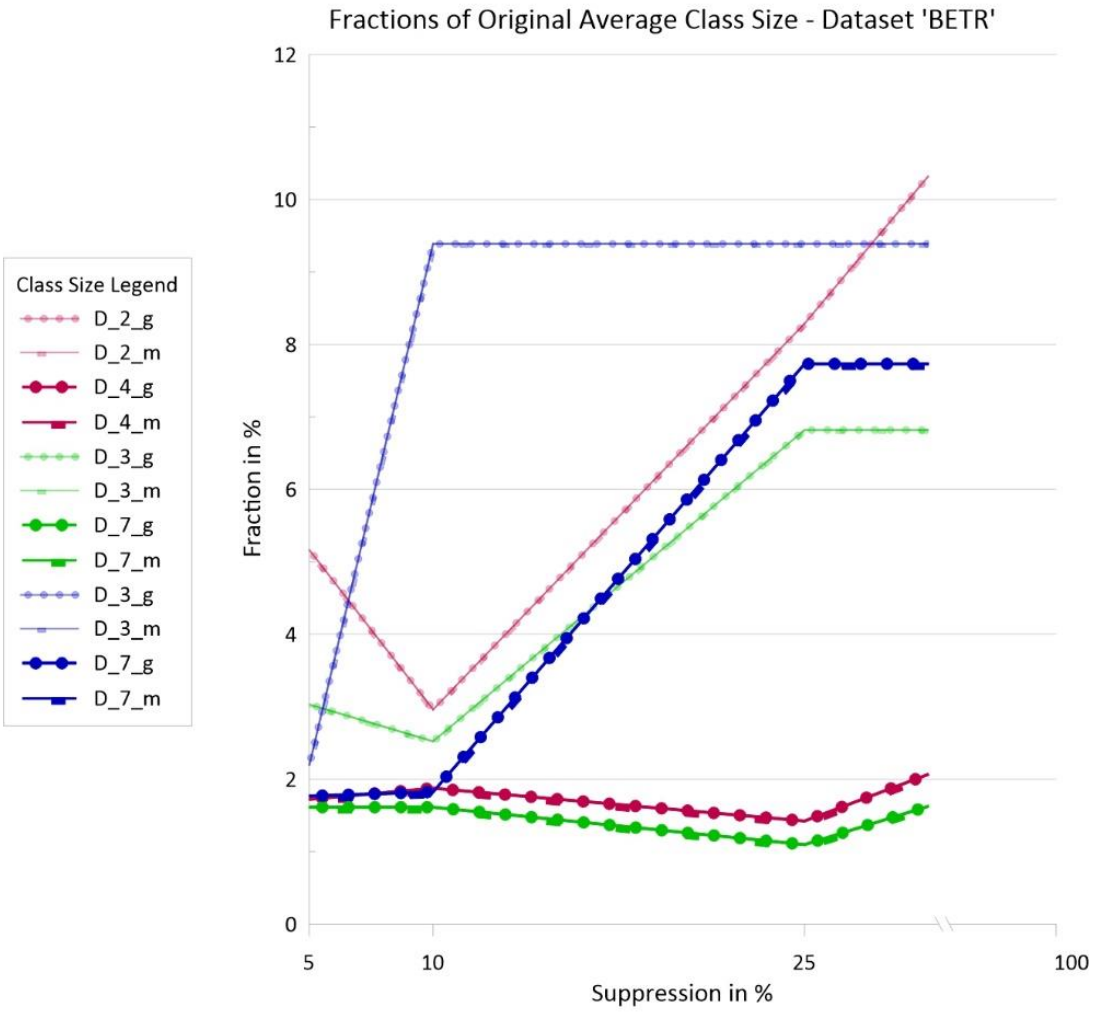
B. Data Table Search Matrix

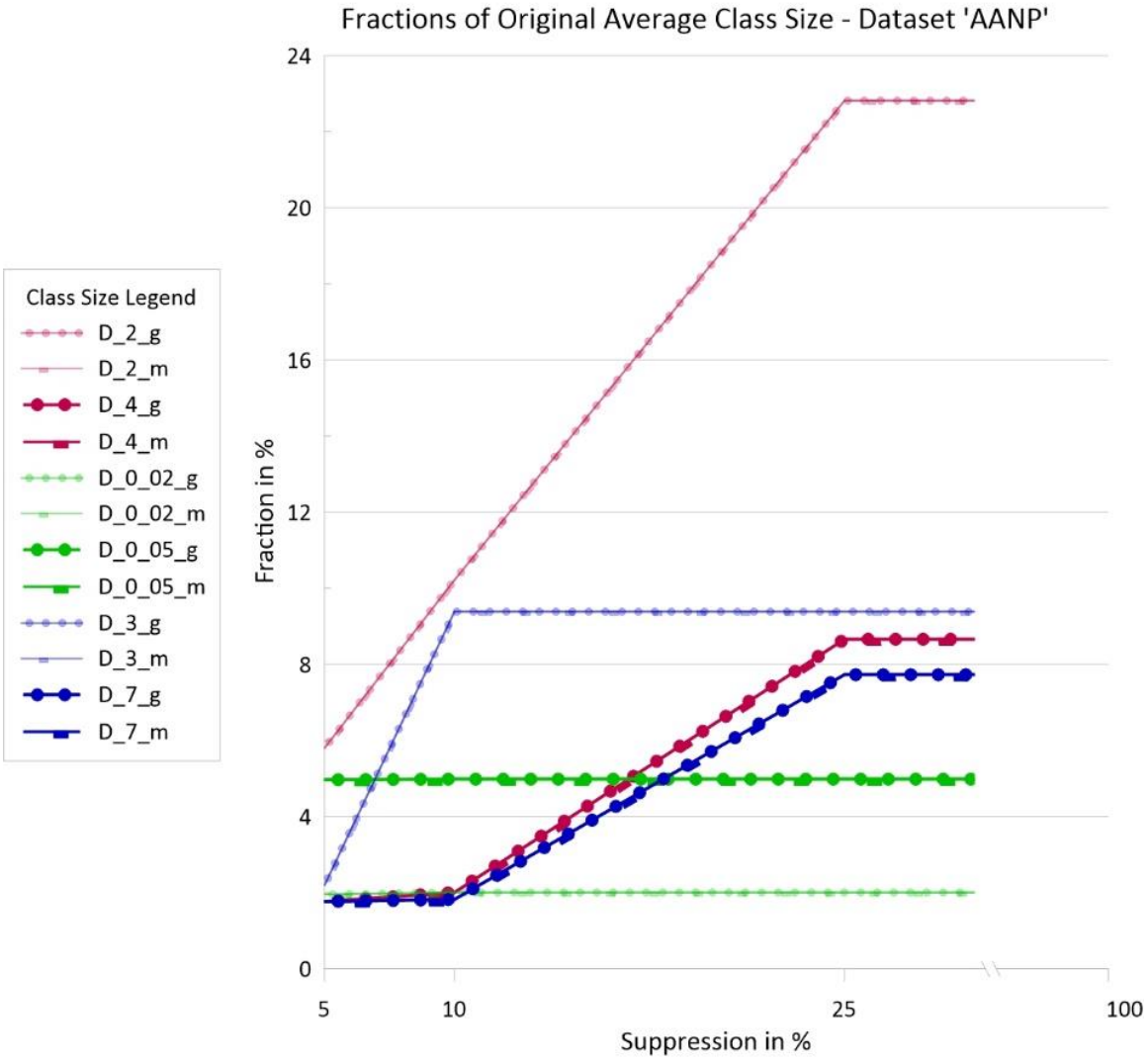
	Client															*Relations*								*Third parties*						
	Identifer		Quasi Idenfier							Sensitive		Semi Sensitive				Identifer	Quasi Idenfier							Identifer	Quasi Idenfier					
	Name	BSN	Adress	Birth date	E-mail	Phone	Status	Lifestyle	Education	Describtion (level 1)	Support	Describtion (level 2)	Support	**Sub**		Name	Adress	Birth date	E-mail	Phone	Status	**Sub**		Name	Adress	Birth date	E-mail	Phone	Status	**Sub**
Agenda_afspraak	1													1																
Agenda_dossierafspraak												1		1																
Agenda_locatie			1											1																
Form_knantwoordstring										1				1																
Form_vraagknoop												1	1	2																
Notificatie_email																											1			1
Regio_aanmelding												1		1																
Regio_abstractgroep																1						1								
Regio_activiteit										1				1																
Regio_adres			1											1																
Regio_bestand												1		1																
Regio_betrokkene_gezin																1				1	1	3								
Regio_betrokkene_standaard																								1			1	1	1	4
Regio_corv_vto												1		1																
Regio_doel											1			1																
Regio_doeltitel										1				1																
Regio_email																											1			1
Regio_factor_optie										1	1			2																
Regio_fo_aanmelding												1		1										1						1
Regio_fo_aanmelding_beoordeling												1		1																
Regio_fo_aanmelding_betreft	1	1	1	1	1	1								6																
Regio_gedragspatroon										1				1																
Regio_ibericht_pgb_toekenning_client			1		1	1								3																
Regio_ibericht_pgb_toekenning_clientrelatie																	1		1	1	1	4								
Regio_ibericht_toewijzing_client	1													1																
Regio_ibericht_ toewijzing_contactgegevens			1		1	1								3																
Regio_ibericht_ toewijzing_relatie																1						1								
Regio_iwmo_adres			1		1	1								3																
Regio_iwmo_client	1	1												2																
Regio_iwmo_realtie																1					1	2								
Regio_leefvorm								1						1																
Regio_logregel_dossier																								1						1
Regio_opleidingsniveau									1					1																
Regio_personengegevens	1	1	1											3																
Regio_problematiekomschrijving										1				1																
Regio_problematiekonderwerp											1			1																
Regio_product										1	1			2																
Regio_product_aud										1	1			2																
Regio_relatiegegevens																1	1	1			1	4								
Regio_risicofactor												1	1	2																
Regio_telefoonnummer						1								1																
Regio_verzoek_om_toewijzing	1	1												2										1					1	2
Regio_werkstatus							1							1																
regio_voorziening_voorziening													1	1																
regio_voorziening_categorie													1	1																
Regio_wijkteam																								1						1
Regio_zorgmail_adres																									1					1
Regio_zorgmail_email			1		1									2																
Regio_zorgmail_medewerker																								1	1			1		3
Regio_zorgmail_patientidentifactie	1	1	1	1										4																

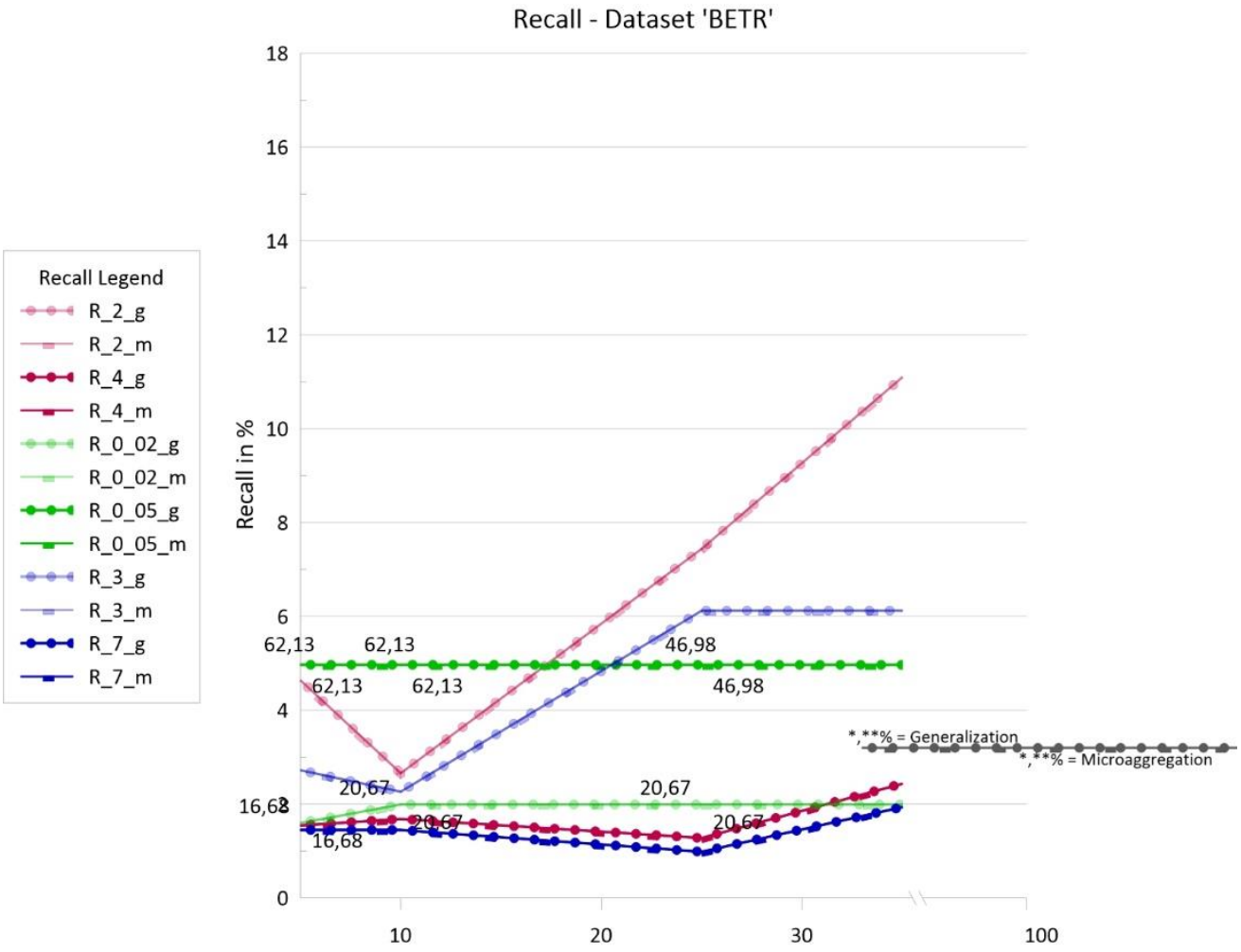
C. Remaining measurements

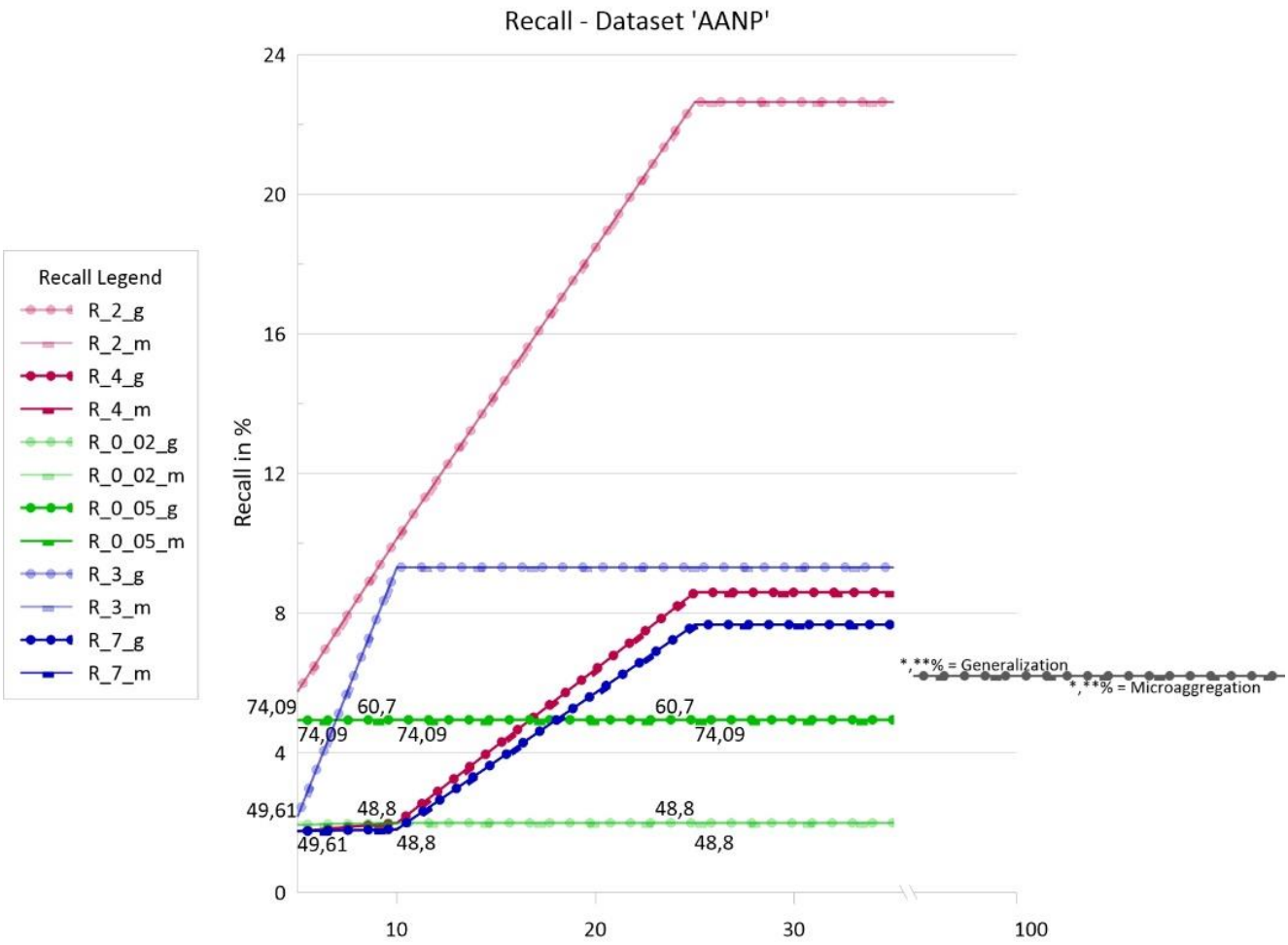


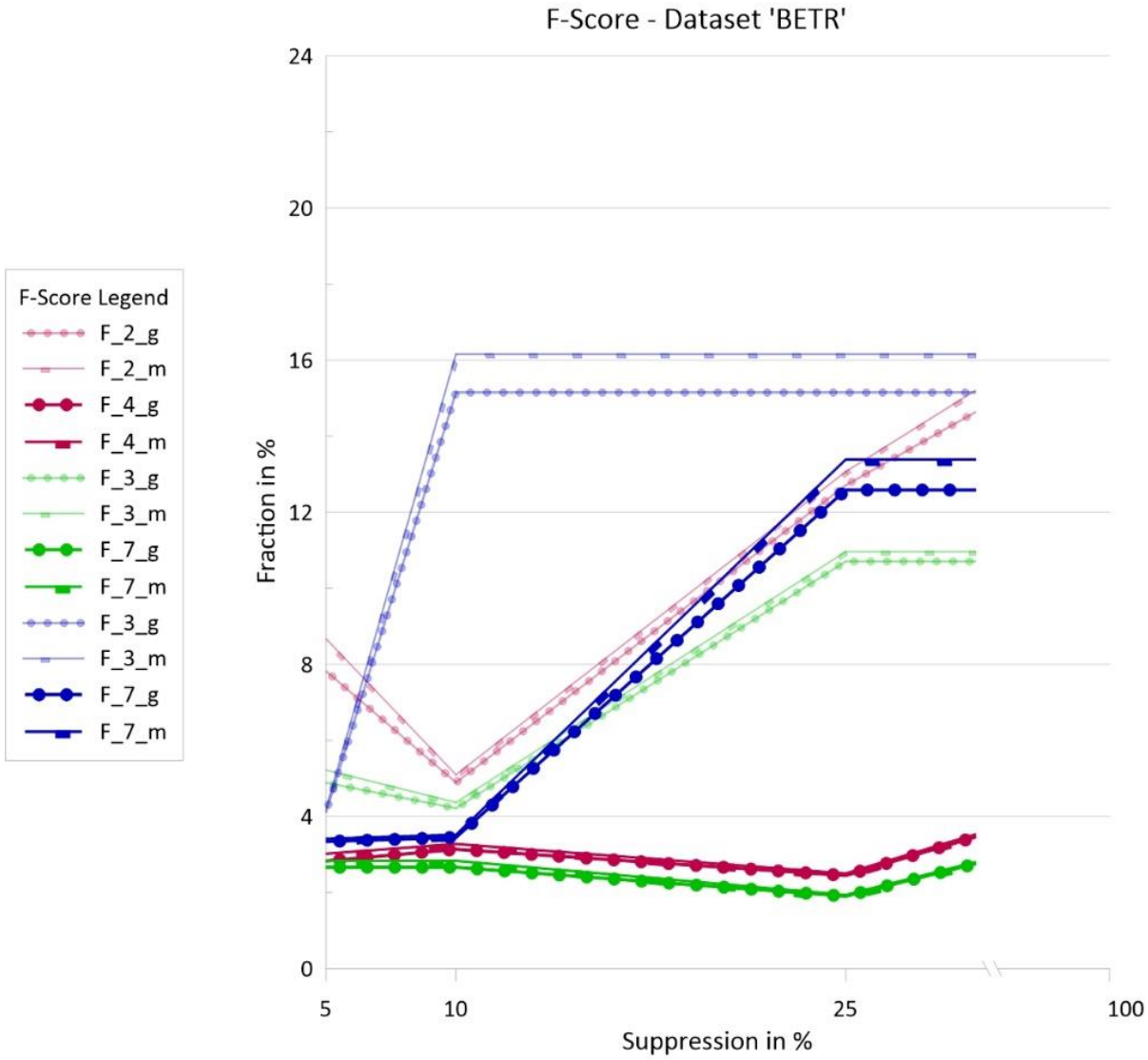

Precision & Record-level Error - Dataset 'AANP'
Precision Legend
G_2_g
G_2_m
G_4_g
G_4_m
G_0_02_g
G_0_02_m
G_0_05_g
G_0_05_m
G_3_g
G_3_m
G_7_g
G_7_m
Error Legend
E_2_g
E_2_m
E_4_g
E_4_m
E_0_02_g
E_0_02_m
E_0_05_g
E_0_05_m
E_3_g
E_3_m
E_7_g
E_7_m
Precision in %
Error Rate in %
Degree of Suppression in %
0
10
20
30
40
50
60
70
80
5
10
25
25
10
5

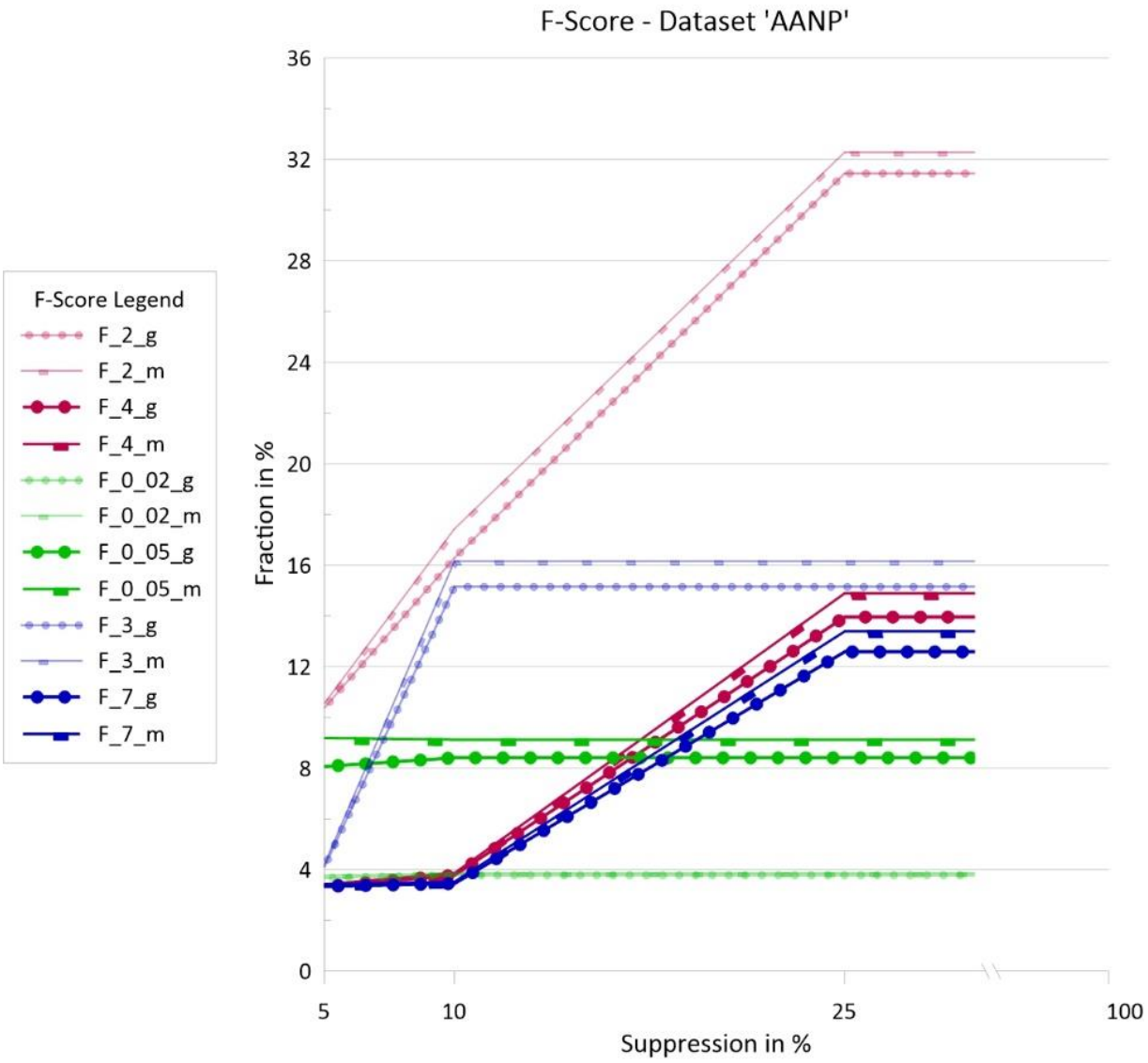

Fractions of Original Average Class Size - Dataset 'BETR'
Class Size Legend
D_2_g
D_2_m
D_4_g
D_4_m
D_3_g
D_3_m
D_7_g
D_7_m
D_3_g
D_3_m
D_7_g
D_7_m
Fraction in %
0
2
4
6
8
10
12
5
10
25
100
Suppression in %


Fractions of Original Average Class Size - Dataset 'AANP'
Class Size Legend
D_2_g
D_2_m
D_4_g
D_4_m
D_0_02_g
D_0_02_m
D_0_05_g
D_0_05_m
D_3_g
D_3_m
D_7_g
D_7_m
Fraction in %
0
4
8
12
16
20
24
5
10
25
100
Suppression in %


Recall - Dataset 'BETR'
Recall Legend
R_2_g
R_2_m
R_4_g
R_4_m
R_0_02_g
R_0_02_m
R_0_05_g
R_0_05_m
R_3_g
R_3_m
R_7_g
R_7_m
Recall in %
18
16
14
12
10
8
6
4
2
0
10
20
30
100
62,13
62,13
62,13
62,13
46,98
46,98
20,67
20,67
20,67
20,67
16,68
16,68
*,**% = Generalization
*,**% = Microaggregation


Recall - Dataset 'AANP'
Recall Legend
R_2_g
R_2_m
R_4_g
R_4_m
R_0_02_g
R_0_02_m
R_0_05_g
R_0_05_m
R_3_g
R_3_m
R_7_g
R_7_m
Recall in %
24
20
16
12
8
4
0
10
20
30
100
74,09
60,7
74,09
74,09
60,7
74,09
49,61
48,8
49,61
48,8
48,8
48,8
*,**% = Generalization
*,**% = Microaggregation


F-Score - Dataset 'BETR'
F-Score Legend
F_2_g
F_2_m
F_4_g
F_4_m
F_3_g
F_3_m
F_7_g
F_7_m
F_3_g
F_3_m
F_7_g
F_7_m
Fraction in %
24
20
16
12
8
4
0
5
10
25
100
Suppression in %


F-Score - Dataset 'AANP'
F-Score Legend
F_2_g
F_2_m
F_4_g
F_4_m
F_0_02_g
F_0_02_m
F_0_05_g
F_0_05_m
F_3_g
F_3_m
F_7_g
F_7_m
Fraction in %
36
32
28
24
20
16
12
8
4
0
5
10
25
100
Suppression in %