

ERASMUS UNIVERSITY ROTTERDAM

Erasmus School of Economic
BSc² Econometrics / Economics Thesis
July 5, 2020

Optimal Aggregation for Marketing Analysis with Privacy Regulations

Supervisor:

P. H. B. F. Franses

Second assessor:

F. J. L. van Maasakkers

Author:

C. Dausend (456088)

Abstract

Nowadays, many companies use personal data to analyze their customers' behavior and implement their findings in marketing strategies. In recent history, the concern about data privacy has intensified due to data breaches. This paper compares different data aggregation approaches to secure individual privacy while simultaneously preserving accurate purchase estimation. Three main strategies are analyzed, namely using *one*, *infinite*, or *multiple* information points. The overall aggregation is performed on a zip code level as this study assumes that people with similar characteristics tend to live in close proximity to each other. The performance of the methods is measured by the coefficient accuracy of simulated data using a hierarchical linear model. The results indicate that aggregating the individual data using one information point, the mean, achieves estimates closest to the true values. In addition, a stable aggregation bias can be observed. Therefore, firms can account for the bias when implementing aggregation to their operations. Finally, the mean approach is tested on an empirical application to food purchase data of the Brazilian e-commerce company Olist.

The views stated in this thesis are those of the author and not necessarily those of the supervisor, second assessor, Erasmus School of Economics, or Erasmus University Rotterdam.

Contents

1	Introduction	1
1.1	Data Collection and Privacy Regulations	1
1.2	Goal and Structure of the Research	3
2	Literature Review	6
2.1	Previous Research on Privacy	6
2.2	Hierarchical Models	9
2.3	Aggregation Techniques	10
3	Methodology	12
3.1	Hierarchical Linear Model	13
3.2	Aggregation Techniques	14
3.3	Performance Measure	15
4	Data Simulation of Organic Meat Purchases	16
5	Simulation Results	18
5.1	Second Layer Coefficient Outcomes	18
5.2	Comparison of Aggregation Techniques	21
6	Real-Life Example of the Mean Technique	23
6.1	Brazilian E-Commerce Olist	23
6.2	Application and Results	24
7	Conclusion and Discussion	25
7.1	Summary and Research Outcome	25
7.2	Limitations, Future Research and Applications	26
	References	28
	Appendix	30

1 Introduction

Recently, online advertisers seem to know your interests and desires before you even know them yourself. For example, imagine you are changing your eating habits, you want to live more sustainable and are trying to buy organic food – probably, you are starting to look for more information, matching recipes, healthy restaurants, or shops online. Normally, you would face a huge variety and volume of products and services, but your browser will most likely automatically present you specific suggestions and ideas that fit exactly your profile.

This phenomenon is called personalized marketing and is possible due to the advanced usage of data analysis and access to a wide range of consumer knowledge, including demographic data, information on interests, location, and more (Smit, Van Noort, & Voorveld, 2014). Is this beneficial or rather scary? Resulting advantages for the customer include a faster way to find desired products with detailed service, informative communication, and sometimes personal discounts, leading to an overall better experience. However, what do you think about privacy security? How much transparency is too much? Consumers are very vulnerable and exposed, which can be dangerous if the data is used for illegal purposes, for example, if data gets shared involuntarily to a third party.

The increasing prospects for data misuse will probably lead to stricter privacy regulations in the future. This paper tries to find an optimal data aggregation technique for marketing analysis that would follow potential privacy laws. Aggregating private data has the benefit that researchers still observe detailed characteristics about specific groups, but are not able to trace it back to an individual. There exists a trade-off between data privacy and estimation accuracy. The goal is to find a method that secures the privacy of individuals, but retains the precise analysis required for personal advertisement to the largest possible extent.

1.1 Data Collection and Privacy Regulations

Today, data collection has become increasingly important for companies to analyze consumers' behaviors and market trends. In marketing research, data collection is defined as the process of gathering and assessing insights on targeted variables. These are then used to answer relevant questions and evaluate outcomes. The term “Big Data“ has been established in 1990 and refers to large data sets that are collected frequently, fast and complex. Big Data is often characterized by four “V’s”: *volume* (large amounts from many different sources), *velocity* (high speed for example through the Internet of Things), *variety* (all types of formats) and *veracity* (reliability and validity)

(McAfee, Brynjolfsson, Davenport, Patil, & Barton, 2012). There are two types of data: primary and secondary data. The former is collected explicitly by the researcher himself to answer the question of interest, and the latter is assembled from other sources, which is often easier and cheaper to obtain (Sapsford & Jupp, 1996). This paper looks at data collection of economic and specific personal information, in particular individual income, which is observed and saved electronically to predict purchases. Such data can be characterized as primary data. However, the methods suggested throughout the paper are going to help firms switch from expensive use of primary data to more secondary data.

To understand the immense increase in data collection, the 2018 report of global data usage from the International Data Corporation (IDC) can be seen in Figure 1. In 2010, a global volume of only 2 zettabytes of data was stored. By 2017, this number increased 13-fold to 26 zettabytes, and is predicted to reach 175 zettabytes by 2025¹.

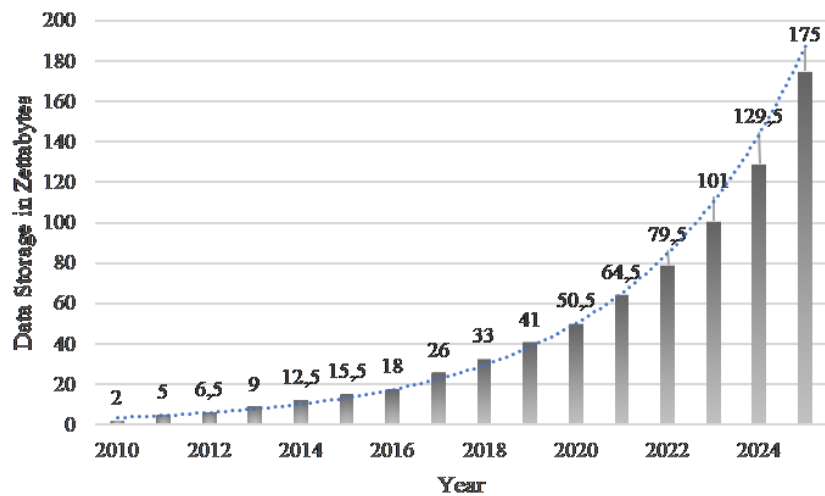


Figure 1: Global Data Storage

Along with this exponential increase, there have been impressive advances in analytics. In the past, people tended to gather data without examining any correlations; it started with simple book-keeping, measuring sales, and first questionnaires. Thereafter, data was used to find explanatory characteristics, which permitted the possibility to forecast consumer preferences. Today there are many techniques to quantify classifications and correlations, like regressions or machine learning analysis. In the new era of technology, many devices like laptops, mobile phones, or smartwatches

¹<https://www.forbes.com/sites/tomcoughlin/2018/11/27/175-zettabytes-by-2025/4183291e5459>

enable companies to get ahold of a lot of personal insights. This leads to very transparent individuals, and methods such as location tracking, eye-scanning, and multi-device fusion are used to understand customer behavior. These systems help to promote the personalization of advertisements or recommendations (Wedel & Kannan, 2016).

Data collection has been growing faster than policymakers could adjust the law. Wedel and Kannan (2016) review different marketing analyses for data-rich environments over the past years and they are among the first to address privacy regulations as a concern from an academic perspective. In May 2018, after almost 20 years without any changes, the European Union passed the new General Data Protection Regulation (GDPR). These rules give citizens more control over their own data. The specific individual freedoms can be specified as follows: the right to be informed, the right to access, the right to rectification, the right to erasure, the right to restrict processing, the right to data portability, the right to object, and also rights around automated decision making and profiling. Moreover, the GDPR deals with seven main principles, including data minimization. For data collection in particular, the guidelines recommend organizations to not collect more personal information than they require for their analysis.² The formulation is vague and challenging to quantify. Thus, it seems probable that the laws will intensify in the future.

1.2 Goal and Structure of the Research

The premises of this research is building on one central assumption, namely that there will be an adjustment of the data protection rules which prohibits firms from storing and analyzing any individual or household data. To still perform marketing analysis, companies could, for example, aggregate data in such a way that only little predictive and diagnostic value gets lost. This leads to the following research question:

“What is the optimal aggregation technique of personal data to secure privacy regulation while minimizing the loss of prediction accuracy of purchases?”

The goal of this research is to provide companies advice on which method they should implement in practice. To tackle this, first, already existing approaches that deal with privacy issues in marketing need to be inspected and evaluated. Subsequently, different aggregation techniques should be discussed and compared. The focus should remain on privacy measures, meaning the methods can not be reversed back to the individual level. The loss of prediction accuracy is tested

²<https://gdpr-info.eu/>

by looking at the regression coefficients. Hence, the closer the estimates to the true values, the more accurate the method. It is of particular interest to investigate in what way the model structure changes and to what extent a difference in coefficients is observed.

In Figure 2, the development of analysis towards data aggregation is presented. Overall, the aggregation is performed on a zip code level, as there is significant evidence that individuals tend to live next to people that share similar characteristics (Steenburgh, Ainslie, & Engebretson, 2003). Three main approaches are considered that rely on *one*, *infinite*, and *multiple* information points. First, one information point per zip code is used for aggregation, namely the mean. Taking the average is known to describe a set of numbers well and is, therefore, often applied in recent literature (Schefter & David, 1985). As a direct comparison, the median is investigated as it could be a better measure in the presence of outliers. One can imagine that including multiple data characteristics might assure less accuracy loss. Therefore, a distribution method is tested where hypothetical households are randomly drawn from a multivariate normal distribution. In practice, data does not necessarily follow a perfectly normal distribution, thus as the last approach, multiple information points are taken to create a discrete distribution aggregation technique (Granger, 1981). Again, individuals are recreated using the zip code quantiles.



Figure 2: Development towards Data Aggregation

All techniques are compared by looking at the second layer coefficients of the hierarchical linear model, which describe the effect of income on the price coefficient. They are evaluated on three simulated examples of monthly purchase behavior of high-quality, organic meat. Each data set contains information about the number of purchases, price, and a household/zip code specific property (income) for a period of one year.

The comparison of the different approaches using various performance measures, shows that the mean technique performs the best and even achieves estimates very close to the values of the unaggregated model. Important to mention is that the evaluation shows a steady aggregation bias. Companies that would implement mean aggregated data could adjust their findings by this given bias, promising very accurate outcomes. A single aggregate is beneficial for companies and

institutions, as taking the average is a lot cheaper and easier to collect and undisputably assures the privacy of its customers. The median outperforms the other methods approximately 10% of the time, but exhibits extreme fluctuations, resulting in significant errors for some scenarios. Therefore, this method is not advised for implementation. The distribution technique produces less precise outcomes than the mean, but interestingly it displays remarkably steady results over all simulations. More research in the corresponding aggregation bias could enhance this method further. Lastly, the quantile technique presents the least favorable results. A possible explanation could be that the relationship between the variables gets lost during aggregation.

As the mean outperforms the other methods, a real-life implementation is provided for this approach. Food purchase data of a Brazilian e-commerce company called Olist is analyzed. The regression outcome are then adjusted by the aggregation bias found in the previous results. This is striking because Brazil published new data regulation (LGPD), which are enforced in August 2020.³ The application should show firms how to implement the technique in their analysis.

Regarding academic relevance, privacy regulations for data protection are still lenient but are becoming stricter. Many studies have focused their attention on using personal information to gain the most accurate and detailed insights, instead of preventing it. This paper attempts to overcome the need for personal data. Furthermore, most research deals with the subject matter by anonymization. A disadvantage is that sometimes de-anonymization methods can retrace the data processes (Narayanan & Shmatikov, 2008). A different approach to privacy modeling compared to previous work is considered. If the companies are not allowed to use personal data, the next logical option is to find smart ways to sum up the information. Therefore, this research compares aggregation techniques. So far, there has been little literature in the field of data aggregation concerning privacy regulation as this topic is ahead of time. As found, simply taking the average works remarkably well without losing a lot of accuracy compared to individual methods, which could have a high impact on future works. Concluding, it aims to contribute to the existing literature by providing a preview of how tightening privacy laws could influence data analysis in the next coming years.

Moreover, contribution beyond the scientific field is achieved. This paper shines a light on the crucial topic of privacy security and correct ways of handling personal data. In doing so, it helps to inform policymakers to make a fairer and more desirable online privacy environment. Many people are not aware of data exploitation, abuse of personal information, or corruption. For example, in

³<https://www.dlapiperdataprotection.com/index.html?t=lawc=BR>

2015, a data breach by AT&T was revealed; employees of the company exposed 280,000 customer information, including names and social security numbers to external parties who then unlocked stolen cell phones for sale on secondary markets.⁴ This study tries to find alternatives to the currently implemented methods, which can not be reverted to the individuals. Social relevance is also achieved by helping companies to perform a smooth transition from their current analysis to aggregate analysis, in case such regulation would be implemented in practice. On the one hand, companies benefit from the findings to secure their marketing strategies and valuable consumer preferences. On the other hand, the public in general is satisfied, as there often exists a so-called privacy paradox, people want their privacy secured, but still like the benefits of personalization (Koorn et al., 2015).

The paper is structured as follows: firstly, the relevant literature is analyzed, which provides the main framework. Subsequently, methods that result from the literature review are explained, and the strategy of providing an answer to the research question is elaborated upon. After this, the simulation of the data set is demonstrated. The methods are adopted, the results are presented, and the different techniques are compared in the following section. A real-life application is fulfilled to apply the outcomes to practice. Finally, the study concludes with the main findings, gives a discussion, list limitations, and suggests further research.

2 Literature Review

There are various ways to predict and model purchase behavior using individual and aggregated data. In this section, a concise review of the existing literature is given. The focus is set on three main aspects: previous research on privacy regulations, purchase estimation with hierarchical models, and different aggregation techniques. A summary table of the most important concepts can be found at the end of section 2.1 and an overview of the methods in section 2.3.

2.1 Previous Research on Privacy

The following paragraphs focus on the implication of privacy and data handling. An extensive summary of existing marketing methods is given by Wedel and Kannan (2016), ranging from simple linear regressions to sophisticated machine learning techniques. The review entails an elaboration of the importance of privacy security and the possible resulting consequences for marketing research. A

⁴<https://www.cnbc.com/2015/04/08/att-data-breaches-revealed-280k-us-customers-exposed.html>

survey by Dupre (2015) mentions that 75% of the respondents believe that firms possess too much private information. Besides the fact that between 2006 and 2016, more than 5000 major data breaches and misuses were revealed leads to the expectation that in the future privacy regulations could tighten. Wedel and Kannan(2016) underline that such laws would make advertisements less efficient, which would particularly harm new and small firms that depend on target marketing. This paper wants to find a method that limits the detrimental impact of regulation on these companies. Therefore, the new aspect of securing household information is tackled through data aggregation. Data aggregation is implemented because it is relatively cheap and easy to obtain and an excellent possibility to be realized by firms in the future (Musalem, Bradlow, & Raju, 2009).

To solve violations against data privacy, one needs to look at the procedure of data analysis. The stages where data misuse is possible need to be located. For that it is important to also investigate the responsibilities of different marketers. This will improve understanding of where it is most convenient to implement new guidelines and alternate the methods to secure privacy. The general process of data mining consists of four main steps which can be seen in Figure 3. Each stage has its own privacy burden (Xu et al., 2014).

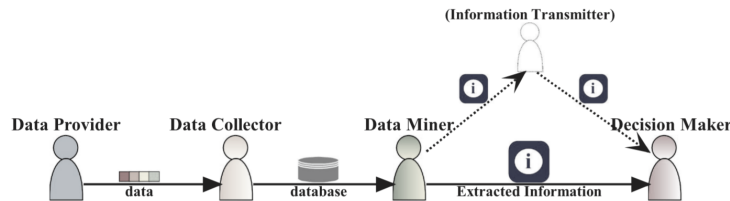


Figure 3: Data Mining Process

In the context of data providers, this paper uses the definition of any individual who releases personal data on the internet. His responsibility is to control the sensitivity of the data he provides to data collectors. As the name suggests, the data collector gathers large amounts of information. He modifies data, so it does not contain sensitive information to ensure data privacy, but still preserves high utility for the data miner. The third stage is data mining, where statistical algorithms are performed to extract useful information. The obligation is to hide sensitive results from untrusted parties. Finally, the decision-maker concludes from the mining results. Therefore to assure privacy protection, he relies on credible results. In this paper, the focus lies on stage two and three, data collection and data mining respectively. Implementing aggregation techniques in these steps would guarantee privacy as single individuals can not be identified anymore(Raghunathan, 2013).

In the past, to prevent privacy concerns, researchers have worked on anonymizing data. Anonymization tries to adjust data across systems so it can not be related to specific individuals, while keeping the format and referential integrity. There are various established approaches, Bayardo and Agrawal (2005) proposed an optimization algorithm for a robust de-identification procedure known as k-anonymization. Data anonymization is a challenging topic as there exists a thin line between not being able to detect the corresponding individuals while still providing detailed information. Often there is a chance of identification. For example, Narayanan and Shmatikov (2008) analyzed a robust de-anonymization technique applied to the media-service provider Netflix. They were able to determine users, including their political preferences and other potentially sensitive information using Netflix’s price records and anonymous movie ratings. This leads to the conclusion that anonymization is not the most appropriate method but aggregation might be a suitable alternative.

Table 1 summarizes the important concepts discussed so far. The theories lead up to a problem that marketing analysts face, namely personalized advertising uses sensitive individual data which creates privacy concerns. This motivates the search for methods that have the protection of individuals as a high priority.

Table 1: Data Collection Concepts

Concept	Description	Author
Personalized Marketing	Companies identify customers’ behavior through data analysis to send individualized messages and product offerings.	Smit, Van Noort, and Voorveld (2014)
Data Collection	Process of gathering and assessing insights on targeted variables, which are used to answer relevant questions and evaluate the outcomes.	Sapsford, and Jupp (1996)
Big Data	Large data collection characterized by volume, velocity, variety, and veracity.	McAfee et al. (2012)
IDC Report 2018	International Data Corporation analyzed the worldwide data storage.	Forbes (2018)
General Data Protection Regulation	European Union law on data protection and privacy.	European Union (2018)
Data Breach	In case confidential information is released to an untrusted environment.	Wedel, and Kannan (2016)
Privacy Paradox	Inconsistency between privacy concern and their desire for personalization.	Koorn et al. (2015)
Data Mining	The process of trying to determine patterns in large data sets, it involves four steps: data providing, data collecting, data mining, decision making.	Xu et al (2014)
Anonymization	Anonymization tries to adjust data across systems so it can not be traced back to a specific individual while keeping the format and referential integrity.	Bayardo, and Agrawal (2005)
De-Anonymization	A reverse data mining method that re-identifies anonymized information.	Narayanan, and Shmatikov (2008)

2.2 Hierarchical Models

In analyzing marketing panel data, there often occurs the problem that data sets include large amounts of units (for example households), but each unit only contains a small number of observations. Such shortage, combined with the goal of finding individual differences, can develop challenges (Wedel & Kamakura, 2012). Frequently observational data follows a hierarchical or clustered structure. For example, individuals within the same geographical area or institutions like schools or companies tend to share similar physical and mental characteristics than individuals chosen at random from the population at large. When clustered data is analyzed via ordinary least squares (OLS) regression, standard errors will be underestimated. They do not take into account the dependency in observation, therefore misrepresenting the statistical significance (McNeish, 2014). Therefore, using simple linear regression might not be efficient.

Hierarchical, also called multilevel models are an advanced form of OLS regression which identify hierarchical structures by allowing for residual components at every level. The model is set up with separate layers for within-unit and across-unit analysis, which are then combined into one single method. The benefit of the hierarchical model is that it treats units as dependent compared to more traditional approaches that analyze the observations separately (Gelman & Hill, 2006). Purchase behavior displays similar patterns, therefore, this study uses a hierarchical model to compare the different aggregation techniques.

This method started to be widely used in the 1980s. Beforehand, there were two types of basic approaches to analyze nested data: disaggregation and aggregation. The former is concerned about the hierarchical data problems by disregarding group deviations. Individual behavior is interpreted without acknowledging the possibility of between-group variation. The latter, on the other hand, ignores personal information. Researchers only observe the group level, therefore, neglecting within-group variation. The hierarchical linear model has the benefit of combining the two levels to a mixed model. It is applied when the explanatory variables are at varying hierarchical levels and have a nested effect on the variable of interest. Besides the advantage of evaluating cross-level data relationships and accounting for shared variances, it also requires the satisfaction of fewer assumptions than other statistical methods (Raudenbush & Bryk, 2002).

An example of a multilevel model estimating purchase behavior is performed by Ruff, Akhund, and Adjoian (2016). The number of fruit and vegetable products bought in a supermarket are analyzed with a particular focus on the effect of a price discount. A hierarchical analysis is completed with a random intercept at the subject level and including a fixed effect at the supermarket level

to adjust for the clustering of individuals within the stores. A random intercept is allowed to vary. This leads to the dependent variable for each individual observation being predicted by an intercept that varies across groups. In this paper, besides choosing a random intercept, varying coefficients are also implemented. It assumes that each group has a different regression model with its own intercept and slope. Including both random elements has the benefit of being very close to reality (Raudenbush, 2004). Furthermore, instead of using personal information, different aggregated data is used. The goal is to get a characterization as accurate as possible.

Steenburgh et al. (2003) analyze marketing campaigns with a combination of individual and aggregated data at the example of the University of Texas. The authors introduce the hierarchical Bayesian variance components model in which they allow a massively categorical variable to be included as an explanatory variable. This variable is used in case there are too many data observations to be treated in a standard manner, so it replicates some of the information contained in the database through aggregation. The authors find promising results, including the aggregate variable, and allows this paper to follow their example.

2.3 Aggregation Techniques

As mentioned above, Steenburgh et al. (2003) use a massively categorical variable, namely, they use zip code level aggregation. It follows from the assumption that people who live in close proximity share similar characteristics. As this hypothesis seems appropriate, this study follows it as well. Three main approaches for aggregation are looked at, including one, infinite, or multiple information points.

One information point: First, the literature including only one data characteristic is inspected. To interpret the zip code features, Steenburgh et al. (2003) are using the average value of their explanatory variables. This is done in many other papers, for example, by Zenor and Srivastava (1993) and seems to be a common approach. Averages are popular because they are known to describe a set of numbers very well. In other words, it is a reliable indicator of what the complete data set would look like. The mean is also a useful measure when the data of interest is spread relatively evenly with few exceptionally high or low values (Pham-Gia & Hung, 2001). In addition, the median is investigated. It is defined as the 50% quantile, thus, the value in a set where half of the numbers are lower, and half of the numbers are higher. The average and median might be very similar, but in case there are significant outliers, the median could represent the group characteristics, such as income within a zip code better (Rosar, 2015).

Infinite information points: One can imagine that using more data characteristics might increase accuracy. Chen and Yang (2007) estimate individuals using aggregate data by performing an augmentation of personal choice. The benefit of the proposed method is that it enables the analysis of micro-level consumer dynamic behavior, for example, the impact of purchase history on current brand choice when only aggregate level data is available. The goal of this approach is to simulate latent choice data that is consistent with the observed aggregate data. The study adopts the idea, hypothetical individuals are recreated by taking random draws of the distributions of the specific zip codes. The aim is to lose as little information as possible compared to individual data. This approach is likely to work well on simulated data, however, in practice, zip codes are less likely to follow normal distributions (Lupton, 1993).

Nowadays, there is a lot of research on recreating individuals in health care. Due to privacy regulations, patients' data is to some extent only available in aggregates. However, to perform proper research about a disease or drug, one needs to investigate on the individual level. One often discussed aspect is to establish the correct relationship between the variables to be able to get concluding results (Wan, Peng, & Li, 2015). If the correlation between the explanatory variables are known or can be approximated, one can select individuals from a multivariate normal distribution. This distribution has the benefit of drawing multiple variables at once, given that they stand in particular relation to each other (Tong, 2012). The downside of this method is that one assumes all variables to be normally distributed.

Multiple information points: Sometimes it is challenging to describe data using a specific distribution. An alternative way is taking a look at the quantiles. Quantiles are points in a distribution that relate to the rank order of values in the distribution (Walker, 1943). For example, the 10th quantile represents the amount where 10% of data is below the value and 90% above, the 50th quantile is equal to the median. Quantiles can help to visualize the distribution of the data. They are typically used in literature to portray certain aspects of a population like income inequality (Bassett, Tam, & Knight, 2003). Here the quantiles are applied to recreate households.

In his book King (1997) addressed the aggregation problem called the ecological inference problem. Here the baseline is that only aggregated data of political votes are available. He wanted to find individual voting behavior with a detailed look at the extend of differences in behavior patterns of race. King (1997) identified a reoccurring problem which is the presence of aggregation bias. To account for that, he focused on the estimation of coefficients of the groups within the aggregation. Related to his findings, this research analyzes the aggregation bias of the four different techniques

and showcase it on an empirical application to actual company data. The bias is estimated by looking at the multiplication factor, which is the value that is multiplied by the estimate to achieve the true value.

In Table 2 a summary of the main methods used through out this paper is given for a better overview. The approach, a short description, as well as the corresponding literature are presented.

Table 2: Methods defined in Literature

Method	Description	Author
Hierarchical Model	Identify hierarchical structures by allowing residual components at every level.	Gelman and Hill (2006)
Fixed vs. Random Effect	Fixed effects are constant across individuals, and random effects vary.	Ruff, Akhund, and Adjoian (2016)
Massively Categorical Variable	A variable, such as zip code, that takes on too many values to treat in the standard manner.	Steenburgh et al. (2003)
Data Aggregation	Process of gathering, summarizing and compiling data.	Zenor and Srivastava (1993)
Zip Code	Geographical area in which people with similar characteristics tend to live.	Steenburgh et al. (2003)
Average	Sum of values divided by the total number of values.	Zenor and Srivastava (1993)
Median	50% quantile, good representation in case of outliers.	Rosar (2015)
Distribution Aggregation	Hypothetical individuals are recreated by taking random draws of the multivariate normal distribution of the specific zip codes.	Lupton (1993)
Quantile Aggregation	Quantiles of zip code variables are used to recreate hypothetical individuals.	Bassett, Tam, and Knight (2003)
Aggregation Bias	The expected difference between the group estimates and the individual estimates.	Gary King (1997)

3 Methodology

This section describes the methods employed to explain and predict the purchase behavior of individuals and aggregates. As can be deduced from the literature, the hierarchical linear model is implemented. It is set up in two layers and looks at the household level data. In the following section, the aggregation techniques are discussed. The three previously described approaches are in compliance with the hierarchical linear model. Lastly, a brief paragraph is devoted to the performance measure that rates the accuracy of different techniques. Below you can find a list of variables that are used in the methods:

Table 3: Variables used in the Models

Index	Variables
Households $i=1,2,\dots,N$	y Quantity of Purchase
Zip codes $z=1,2,\dots,Z$	x Price
Time $t=1,2,\dots,T$	w Property of Household (Income)

3.1 Hierarchical Linear Model

The hierarchical linear model is applied because the effect of explanatory variables are expected to vary over hierarchical levels and have a nested effect on the variable of interest. Therefore a two-level linear model is implemented. The dependent variable is set as the average number of purchases of a product per month. It is assumed that product information and purchase behavior influence the dependent variable, which is represented in Equation (1), namely the number of purchases per household is explained by the price and previous purchases as explanatory variables. Besides, household characteristics can determine the purchase behavior through the impact of the previously mentioned variables. Accordingly, in the second layer, the significance of household specific characteristics on price and previous purchase effects are analyzed. In Equation (2) the price coefficient β_i is regressed on income, one could imagine people earning high salaries might be willing to spend more or less money for a certain product than individuals with a lower income. For the scope of this paper, the results are concentrated on finding an aggregation technique that describes β_0 and β_1 most accurately. The hierarchical model with its two layers is displayed below. First layer:

$$y_{i,t} = \alpha_i + \beta_i x_{i,t} + \phi_i y_{i,t-1} + \epsilon_i. \quad (1)$$

Second layer:

$$\beta_i = \beta_0 + \beta_1 w_i + v_i, \quad (2)$$

The two-level hierarchical model requires the estimation of different types of parameters. Fixed effects do not change across groups (Hofmann, 1997). Here the fixed effects are defined by β_0 , and β_1 in Equation (2). The second type of parameters are the random coefficients α_i , β_i and ϕ_i , which can vary across groups. To estimate the coefficients ordinary least square regressions are applied.

3.2 Aggregation Techniques

Next, the aggregation techniques are explained. In total there are the three different approaches (one, infinite and multiple information points) that are performed. The zip code is used as the classifier on which the aggregation is executed. The household (i) characteristics change to properties of zip codes (z). Therefore the model develops into the formulation below (Equations (3) and (4)): First layer:

$$y_{z,t} = \alpha_z + \beta_z x_{z,t} + \phi_z y_{z,t-1} + \epsilon_z. \quad (3)$$

Second layer:

$$\beta_z = \beta_0 + \beta_1 w_z + v_z, \quad (4)$$

Note that the model structure barely changes, only the input data. The changes in the variables are discussed below. In the following, the term X_z is used to represent all variables, including $y_{z,t}$, $y_{z,t-1}$, $x_{z,t}$, w_z .

One information point:

Mean Technique: The method carries out the average per zip code (Equation (5)), which is calculated for all explanatory variables and the dependent variable as they all contain individual elements. The mean is the most intuitive and is widely used in practice. One needs to take into account that for the income the aggregate is constant over time, whereas the aggregate varies for purchases and price:

$$\bar{X}_z = \frac{1}{N} \sum_{i=1}^N X_i. \quad (5)$$

Median Technique: As a direct comparison, the median is investigated (Equation (6)). The median represents the 50% quantile. This method is interesting because the median could protect estimation results from influences of outliers:

$$X_{z.median} = \begin{cases} X_{\frac{N+1}{2}} & \text{if } N \text{ odd,} \\ \frac{1}{2}(x_{\frac{N}{2}} + X_{\frac{N}{2}+1}) & \text{if } N \text{ even.} \end{cases} \quad (6)$$

Infinite information points:

Distribution Technique: The following strategy uses the features of the zip code distributions. It is assumed that every zip code variable (X_z) follows a specific distribution, for example, income usually follows a log-normal distribution (Clementi & Gallegati, 2005). The method tries to recreate household characteristics. As such, every observation is drawn randomly from a multivariate normal

distribution, denoted \tilde{i} (Equation (7)). This has the advantage of securing the relationship of variables by calculating the covariance matrix (Σ), but it entails the assumption that all variables are normally distributed within each zip code:

$$X_{\tilde{i}} \sim N(\mu, \Sigma), \quad (7)$$

where

$$\mu = \begin{bmatrix} \bar{X}_1 \\ \dots \\ \bar{X}_z \\ \dots \\ \bar{X}_Z \end{bmatrix}, \Sigma = \begin{pmatrix} \sigma_1^2 & \dots & \sigma_{1Z} \\ \dots & \dots & \dots \\ \sigma_{Z1} & \dots & \sigma_Z^2 \end{pmatrix}. \quad (8)$$

Multiple information point:

Quantile Technique: Most of the time, in practice, data does not perfectly follow a distribution. Therefore, the last method tries to recreate hypothetical households again by looking at the zip code quantiles instead. The 10th, 20th, 30th, 40th, 50th, 60th, 70th, 80th, 90th quantiles are calculated. So for every quantile you double the observations until you have N observations per quantile, where N is the total number of households per zip code.

3.3 Performance Measure

To answer the research question, the optimal aggregation technique is evaluated by considering five different performance measures (Equations (9) to (12)). The benefit of applying the aggregation techniques on simulated data is that the true coefficients are known. The closer the estimates are to the original values, the more accurate the data description. The parameters of interest are the regression coefficients, in particular the factors in the second layer, namely β_0 and β_1 . In the following β represents the true value and $\tilde{\beta}$ the aggregation estimate.

First of all, the **average coefficients** ($\bar{\beta}$) over all simulation runs is compared to the true values (Equation (9)). This measure gives an intuition of the general performance of the estimates. Here, it might also be interesting to look at the standard deviation of the coefficients, the smaller the deviation, the more stable the coefficient estimates over the different simulation runs:

$$\bar{\beta} = \frac{1}{R} \sum_{r=1}^R \tilde{\beta}_r. \quad (9)$$

Using the average values, the **aggregation bias** (AB) is also calculated (Equation (10)). It is of interest if the bias stays the same over the different coefficient specifications:

$$AB = \beta/\tilde{\beta}. \quad (10)$$

Also, the **coefficient development** over all simulation runs is looked at. Hence not only the distance of the estimates to the true values is of importance but the less fluctuations can be seen, the better.

Thereafter, the estimations are compared by looking at the **relative error** (RE), (Equation (11)). It can be calculated by dividing the absolute error by the magnitude of the exact value. This is interesting because the measure gives an inferior assessment of the severity of the estimation as opposed to considering the absolute difference:

$$RE = \frac{\beta - \tilde{\beta}}{\beta}. \quad (11)$$

Finally, a **hit rate** table is constructed, to study which method (m) performs best overall (Equation (12)). The hit rate is defined as the percentage of times a certain method was closest to the true value compared to the other aggregation techniques:








$$HitRate_m = \frac{\# \text{ of minimal differences for m}}{\# \text{ of runs}}. \quad (12)$$

4 Data Simulation of Organic Meat Purchases

To be able to analyze the aggregation techniques mentioned above, purchase data is simulated using product and household characteristics. As an example, monthly organic meat sales are imitated. Important to note is that in total three different simulations are performed using pre-specified coefficient scenarios. This is done to investigate if the aggregation methods vary in their performance for different values.

Each simulation is set up using the two-level hierarchical linear model like described in the methodology. First, the second layer explanatory variables are randomly drawn from a distribution, and the coefficients are set. From these values, the factors in the first layers are calculated. Finally, the explanatory variables from the first layer are also randomly drawn. The dependent variable is then computed using all the previously established variables and coefficients. This procedure is performed 1000 times, the resulting data sets are used to estimate the aggregation outcomes. Table 4 shows the specifications for the main variables and coefficients used for this research.

Table 4: Descriptive Statistics for Simulation

		Value		
	Simulation Runs	1000		
	Time Period	12 Months		
	Zip Codes	50		
	Households	100 x 50 zip codes		
	Purchases at Time 0	10		
		Distribution	Mean	St. Dev.
	Income	Log-Normal	35	3
	Price	Normal	10	1
	ϕ_i	Normal	0.4	0.1
		Simulation 1	Simulation 2	Simulation 3
	β_0	-1	-2	-3
	β_1	0.04	0.06	0.10

The data sets are constructed as follows: there are 50 groups which represent zip codes, each group contains 100 households. The dependent variable represents the average household purchase of organic meat per month for a total period of one year. Summing up, one data set contains 60,000 household and 600 zip code observations.

The **second layer** is modeled first. As can be seen in Equation (2), the explanatory variable is household income. Note that the income stays constant over all time periods. To account for the assumption that people living in the same zip code tend to have similar characteristics, separate draws are done for each of the 50 zip codes. Clementi and Gallegati (2005) analyzed that income typically follows a log-normal distribution. Therefore, the average income per zipcode is drawn randomly from a log-normal distribution with mean of €35,000 a year and standard deviation of €3,000, which is according to the dutch average in 2018.⁵ The values are taken in thousands, therefore 35 and 3 accordingly. Every draw corresponds to a specific group and act as the average income per zip code. Within the zip code the household incomes are then picked randomly from a

⁵<https://www.statista.com/statistics/538406/average-annual-salary-in-the-netherlands-by-age/>

normal distribution using the before mentioned average income and standard deviation of €1000, so 1 respectively.

As was just mentioned three simulations with different coefficient scenarios are performed. The only difference across the simulations are the specification of the second layer, which are set by an educated guess. β_0 is defined as a negative number, -1, -2, and -3 because the average price effect is negative, and β_1 is fixed small, 0.04, 0.06, and 0.1, as the income values can get high. The given quantities are then used to calculate β_i . ϕ_i is directly drawn from a normal distribution with a mean of 0.4 and a standard deviation of 0.1.

The **first layer** of the hierarchical linear model is set up next. Price and previous purchases are taken as explanatory variables. The price variable for one portion of meat per household i at month t is generated randomly from a normal distribution with a average price of €10.00 and standard deviation of €1.00. These numbers were approximated from the prices of organic beef steak sold at Ecoplaza in the Netherlands.⁶ To calculate the previous month's purchases, the average amount is set to 10 portions for every household at time 0. For the following months, the variable is then calculated using Equation (1) and inserting the coefficients and prices.

The zip code statistics such as average, median, standard deviation, and quantiles are determined from the resulting individuals belonging to the specific group, which are then used to perform the different aggregation techniques.

5 Simulation Results

This section first provides the results found from performing the hierarchical linear model on the simulation data using the different aggregation techniques. Then the methods are compared and the optimal aggregation approach is evaluated.

5.1 Second Layer Coefficient Outcomes

Performing each simulation 1,000 times results also in 1,000 estimates for each coefficient. As a first performance measure, the average β_0 and β_1 over all simulation runs are looked at. In Table 5, the outcomes are shown for every aggregation technique as well as for every coefficient scenario. One can observe that the mean approach generates estimates close to the true values. The distribution and the median estimates are less precise. The quantile approach performs considerably worse.

⁶<https://www.ekoplaza.nl/producten/product/runder-biefstuk>

Interestingly, only the mean slightly overestimates the coefficient values in absolute terms, whereas all the other techniques underestimate. The standard deviation is very small for the distribution technique. The mean and quantile approach show rather limited deviation as well, which means the results do not vary that much and evolve around the mean. In contrast, the median range it is slightly bigger. The standard deviation of the coefficients are very similar across the various scenarios for all the aggregation techniques.

Table 5: Estimation Evaluation Comparison

	Household Level		Zip Code Level			
	True Values	Estimated values	Mean	Median	Distribution	Quantile
β_0	-1	-0.9993 (0.02164)	-1.0070 (0.2217)	-0.6869 (0.5597)	-0.6361 (0.0478)	-0.6806 (0.2045)
	-2	-1.9993 (0.0218)	-2.0007 (0.2258)	-1.2787 (0.5974)	-1.4572 (0.0636)	-1.1622 (0.2055)
	-3	-2.9993 (0.0217)	-3.0085 (0.2388)	-1.9923 (0.8423)	-2.0911 (0.0852)	-1.7760 (0.2939)
β_1	0.04	0.0310 (0.0007)	0.0402 (0.0063)	0.0270 (0.0161)	0.0297 (0.0014)	0.0261 (0.0059)
	0.06	0.05998 (0.0007)	0.0602 (0.0064)	0.0384 (0.0171)	0.0445 (0.0018)	0.0351 (0.0059)
	0.10	0.1000 (0.0007)	0.1003 (0.0068)	0.0663 (0.0246)	0.0742 (0.0025)	0.0593 (0.0087)

As King (1997) explained, one has to account for the aggregation bias. Therefore, the multiplication factor is analyzed. The multiplication factor is the number that needs to be multiplied with the estimation outcome to achieve an inferior estimate, as it is empirically closer to the true value. The factors are presented in Table 6. One can see that the mean stays steady, with an average adjustment of 0.9966 for β_0 and 0.9962 for β_1 . Note, there is a slight upward trend when the β_1 increases. For the other methods, the outcome is not that direct, the values vary more, and one can not see a trend as the coefficients increase. Only the bias for β_1 for the distribution technique stays steady as well.

Table 6: Aggregation Bias

	β_0				β_1			
	-1	-2	-3	Average	0.04	0.06	0.1	Average
Mean	0.9930	0.9997	0.9971	0.9966	0.9950	0.9966	0.9970	0.9962
Median	1.4558	1.5640	1.5058	1.5085	1.4815	1.5625	1.5083	1.5174
Distribution	1.5721	1.3725	1.4346	1.4597	1.3468	1.3484	1.3477	1.34676
Quantile	1.4694	1.7209	1.6891	1.6264	1.5326	1.7094	1.6863	1.6428

To get an indication of how the coefficients change across the simulation runs, β_0 is plotted in Figure 4 and β_1 in Figure 5 for the first 100 simulation runs. The full development can be seen in the Appendix. For both parameters, the mean evolves around the true value but shows some volatility. The median entails substantial movement and outliers. It has values far below and above the true line, but also touches it sometimes. The distribution estimation stay rather constant and moves around the average value given in Table 5. The distribution lays above β_0 , and below β_1 , it never touches the true value line. Finally, the quantile approach is comparable to the distribution line but includes more spices. Similar findings can be seen in the full simulation graphs in the Appendix.

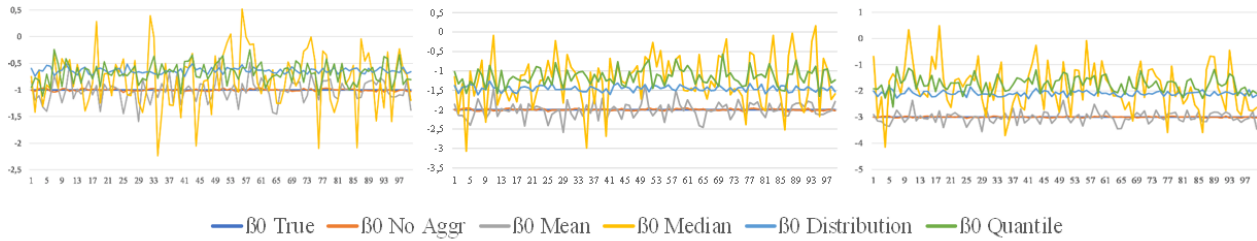


Figure 4: β_0 Development over the first 100 Simulation Runs

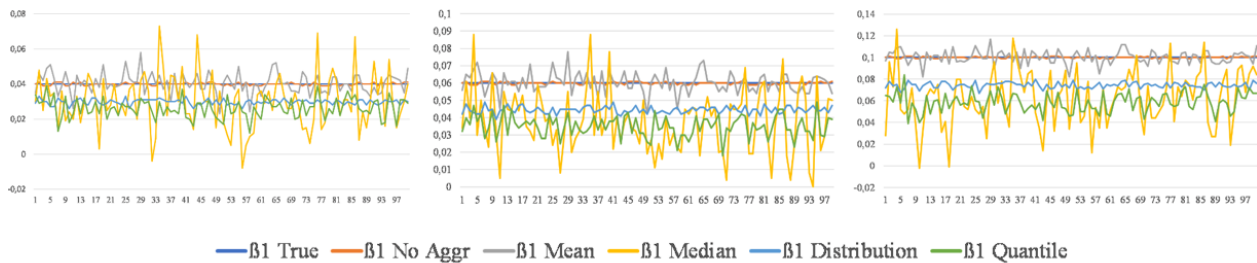


Figure 5: β_1 Development over the first 100 Simulation Runs

The relative error gives the percentage difference between the estimated and the true value. In Table 7, the average relative errors are given, the coefficient specific relative errors are in the Appendix. These values are particularly interesting as they provide insights into the severity of the estimation mistakes. For the mean, the RE is relatively small, close to -2% and -1% for the respective coefficients. This is somewhat unexpected, as in Table 5 one can see that the mean overestimates, so it would have been intuitive to get a positive RE. Note that for β_0 equal to -1 and β_1 equal to 0.04, the error is slightly more negative than for the other two values. The median approach shows a remarkably large negative error. It performs better for β_0 equal to -1 and β_0 to 0.04, so the opposite of the mean. The distribution technique shows a relative small RE. This is to be explained by the minor deviation of the average value and the limited frequency of the presence of outliers, as depicted in Figure 4 and 5. The quantile regression provides a rather poor RE for both coefficients. Note that in general the RE is slightly less severe for β_1 than for β_0 .

Table 7: Relative Error Comparison

	Household Level	Zip Code Level			
	Estimated values	Mean	Median	Distribution	Quantile
β_0	-0.0633%	-2.2075%	-157.8842%	-46.4460%	-72.7592%
β_1	-0.0537%	-1.1053%	-78.5213%	-34.8980%	-70.5129%

5.2 Comparison of Aggregation Techniques

The following paragraphs compare the different aggregation methods. As could be seen in Table 5 the mean estimate resembles the true value the most as opposed to the other methods. The distribution and median follow and the least accurate average estimates are given by the quantile regression. Interestingly, only the mean slightly overestimates the coefficient values in absolute terms while all the other techniques underestimate.

From the average values it follows that also the mean shows the smallest aggregation bias. Another favorable aspect is that the bias stays very constant. The variation in the bias created by other methods are considerably larger and vary for every coefficient scenario. The distribution method achieves second smallest bias, followed by the median and then the quantile estimation.

The coefficient development shows that the mean moves around the true value, but has more fluctuations than the distribution technique. The distribution shows by far the least variations,

however simultaneously is notably distant from the true value. The quantile regression performs similar to the distribution method in terms of the distance to the true value, but fluctuates fairly. The median realizes the worst development with extreme outliers.

As the last performance measure, the hit rate is considered. The hit rate represents the percentage of time a specific estimation method outperforms the other methods. Therefore this measure is an essential evaluation to determine the optimal aggregation technique. Below in Table 8, the hit rates over all simulation runs, including all three coefficient specifications, are represented. Scenario specific outcomes can be found in the Appendix. The mean zip code aggregation is closest to the true value in 84.3% of the times for β_0 and 84.1% for β_1 , which means it aggregates the data most effectively. The median follows with roughly 10% for both coefficients. This outcome is to some extent surprising because the median had a worse average, RE and coefficient development than the distribution technique. The distribution technique only achieves a hit rate for approximately 2% to 3% for either coefficient. Finally, the quantile approach never comes closest to the true β_0 compared to the other techniques, but in about 4% of the time it is most accurate for β_1 .

Table 8: Hit Rate Table

	β_0	β_1
Mean	84.3%	84.1%
Median	12.7%	9.6%
Distribution	2.9%	2.2%
Quantile	0%	4.2%

Summarizing, one can observe that the mean estimation outperforms the other methods, achieving close results to the true value. Furthermore, the median seems to achieve accurate results for some data set as well, however, due to substantial outliers in the coefficient development, it can be concluded that the technique is unstable and might not result in reliable estimates. The distribution approach provides rather erroneous estimations compared to the true values but shows a very stable evolution. Lastly the quantile approach performed remarkably poor in all five performance techniques as opposed to the other techniques.







6 Real-Life Example of the Mean Technique

6.1 Brazilian E-Commerce Olist

In the previous section, one could observe that taking the mean as an aggregation technique performs best and yields estimates that are very close to the individual level analysis. Therefore, in this application, the mean is applied to real-life data of a Brazilian e-commerce company called Olist.⁷ A Brazilian company is chosen, as Brazil enacted the Brazilian General Data Protection Law (LGPD). These regulations were published in 2018 and they closely follow the European GDPR. Most of the laws are going to be enforced in August 2020.⁸ With the help of this paper, companies can prepare for their future marketing transformation towards more secure data analysis and learn how to implement aggregation methods in their operations. Furthermore Olist is chosen as it provides data on food purchases which is in line with the simulation example.

In Table 9 a summary of descriptive statistics of the data, including purchases, price, and income can be seen.

Table 9: Olist descriptive statistics

		Value	
	Time Period	8 Quarters (2017-2018)	
	States	15	
	Purchase at Time 0	1	
		Mean	St. Dev.
	Purchases	0.7986	0.5494
	Price	63.2188	35.9864
	Income	1230.1030	392.3087

The whole data set includes information of 100,000 orders, however, in this application the focus is only on food purchases for the period 2017 to 2018 made at multiple marketplaces in Brazil. Four regions provide less than 5 observations and are excluded from the data set, resulting in a total of 774 observations. The average is taken on the state level due to limited data availability at the zip

⁷https://www.kaggle.com/olistbr/brazilian-ecommerce?select=olist_geolocation_dataset.csv

⁸<https://www.dlapiperdataprotection.com/index.html?t=lawc=BR>

code level, but the assumption of similarities in characteristics within states is not to be classified as unreasonable. Instead of looking at monthly data, the information is given at a quarterly frequency. For time periods where there are no items purchased, the quantity is set to zero, and the average price over all periods is taken. The mean income per state is used from ipeadata, the series is calculated from the responses of the National Household Sample Survey.(Pnad / IBGE) from 2014⁹

6.2 Application and Results

To take into account that the quantities range between 1 and 2, but the price and income take on large numbers, the price is adjusted by dividing by 10 and income by 100, for visualization purposes of the coefficient effects. Then the data is aggregated to the mean state values. The regression is performed using a hierarchical linear model, as explained in the methodology. The result section showed that there is an aggregation bias of approximately 0.005% when using the mean approach, therefore, it is advised to adjust the regression result by that amount. Below the second layer coefficients are represented as well as the adjusted values.

Table 10: Olist descriptive statistics

	Regression Outcome	Multiplication Factor	Adjusted Regression Outcome
β_0	0.0419	x 0.9966	0.0418
β_1	-0.0099	x 0.9962	-0.0098

Given Table 10, the company Olist would know that an increase in income would lead to a small decrease of the effect price has on food purchases. The conditional mean of the price effect on food purchases is slightly lower. The benefit of a one information aggregation is that it barely changes the analysis model and the information is simpler and cheaper to obtain. Therefore it would be beneficial for Olist to implement the mean aggregation method as it would assure data privacy, give accurate estimations and it would safe time and cost that inherently arise from the collection of data.

⁹<http://www.ipeadata.gov.br/Default.aspx>

7 Conclusion and Discussion

Last but not least, the paper finishes with a conclusion by summing up the most important results and findings to give a concise answer to the research question. Then, a discussion follows, stating the limitations, suggestions for further research and possible applications.

7.1 Summary and Research Outcome

Nowadays many companies use personal data to analyze their customer behavior and implement their findings in marketing strategies. In recent history, there has been more concern about data privacy due to data breaches and misuses. This paper aims to compare different data aggregation approaches to secure individual privacy but at the same time assure accurate purchase estimation for companies. At the moment there are no laws that restrict data collection to an aggregated level, however this research assumes such restriction in the future. Therefore, this study intends to provide a preview on how data analysis would change in case tighter privacy laws would be implemented. Three main approaches are analyzed: using *one*, *infinite* or *multiple* information points. The aggregation is performed on a zip code level because people with similar characteristics tend to live in close proximity to each other. The performance of the methods is measured by the coefficient accuracy of simulated data using a hierarchical linear model.

The fundamental result of this paper is that among all models, using one information point performs the best, namely the mean aggregation. In all five performance measures the approach reaches outstanding results. Not only does it outperform the other aggregation techniques but it also achieves estimations close to the true value. Another advantage is that there is a clear bias tendency which can be incorporated when applying this method. This can be seen in the real-life example of Olist. Another finding is that the median performs well when referring to the hit rate as it outperforms the other methods in about 10% of the time. This is to some extent surprising because the average value over all simulation runs shows a significant difference to the true values. There are large fluctuations, which cause the results to vary a lot, leading to the conclusion that this method should not be implemented to real-life applications. Furthermore the distribution method shows rather unreliable results compared to the mean, as the average estimation is considerably different than the true value. Note, that the distribution constructs the most stable results over the simulation runs. The quantile method displays fairly poorly outcomes. It should therefore not be considered in practice.

In summary, to answer the research question of which aggregation techniques is optimal to secure privacy regulation without losing prediction accuracy, one can conclude that aggregating personal data to one information point, in particular aggregating to the mean, results in the best and most accurate estimation. The difference in coefficient value is small and can even be adjusted by an aggregation bias to reach the true factor. Another advantage it that the model structure barely changes, therefore companies could easily implement this to their operations.

Wedel and Kannan (2016) stated that following stricter privacy regulations would lead to less efficient marketing analysis, which could hurt in particular new small firms that depend on target advertisements. However, with this paper it is proven that aggregated estimation, in other words privacy secured estimation, loses only little accuracy and, therefore, has sufficient predictive power. Aggregation of data is cheaper and more accessible to get a hold off, as oftentimes one can make use of secondary data instead of primary data. Concluding that marketing analysis might end up being more efficient than prior.

7.2 Limitations, Future Research and Applications

When one attempts to compare aggregation techniques using simulated data, there are some biases and limitations that need to be considered. The next paragraphs list these restraints and present various extensions for future research.

First of all, most of the data is set freely or drawn randomly from a normal distribution. In practice data usually does not tend to perfectly follow such a distribution. The simulated data therefore shows symmetric processes, which could lead to the mean and median to perform remarkably better results than in practise. Also, the distribution technique discussed might yield worse results as it draws individuals from a multivariate normal distribution. It would be appealing to apply the different techniques to actual company data where one can compare individual level estimation with the aggregated methods. Another point of interest would be to investigate the distribution technique for different distributions other than the normal distribution. Because of the fact that the distribution outcomes are steady across the simulation runs, it leads to the speculation that researching the aggregation bias more extensively could develop into very accurate results.

This paper has taken into account fixed and random effects to provide more insights and model the hierarchical data structure. But dealing with panel data means information is collected over time and over the same individuals. The regression is run over these two dimensions, and even though this is touched upon in this research it could be improved further in the future.

In addition the aggregation techniques are performed on a hierarchical linear model. As all conclusions are related to this model, it could be the case that there are variations in the aggregation performance when implemented in different models. Another limitation is the use of OLS to estimate the coefficients of the separate layers. As the explanatory variables are correlated with the error term, one finds endogeneity, which is a violation of the OLS assumptions. However, this paper's aim was to find the optimal aggregation technique, not the best estimation model. The use of the hierarchical models often disregards the assumption of homoscedasticity, as the error terms are likely to vary across the independent variables. In particular, one could observe heteroscedasticity in the second layer of the hierarchical model. To resolve this problem, one could for example implement generalized least squares regression in the second layer. Therefore, further research could be done by taking this assumption into account and analyzing its effect on the aggregation methods.

The real life implication is presented to give companies an idea of how they can implement the aggregation in their operations. The company Olist was chosen as there are currently data privacy enforcement's happening in Brazil. Due to data availability the data set however only contained roughly 750 observations, which is enough to give the aggregation example but not sufficient to draw clear conclusion from it. As this paper is addressing marketing analysis in general, including big data analysis, it would be interesting to investigate how a real-life example with extensive observations would perform.

Moreover, the paper solely focuses on the estimation of purchases. There are many areas where personal data is used. Accordingly, it would be interesting to investigate if the aggregation methods would be a good possibility to implement in other fields as well. For example analyzing financial services of banks, researching disease developments in healthcare or policy making in governments.

This paper has added valuable insights into aggregation techniques. It confirms that papers which implemented the mean approach are using an accurate technique. It gives a legitimate alternative to the anonymization techniques which are researched so far. Furthermore, with the real life example a preview of how tightening privacy laws would influence data analysis in the coming years. Concluding, the intention of this paper is to help companies perform a smooth transition to data aggregation, to preserve data privacy and to keep marketing analysis as personal as possible for marketing.

References

- Bassett, G. W., Tam, M.-Y., & Knight, K. (2003). Quantile models and estimators for data analysis. In *Developments in robust statistics* (pp. 77–87). Springer.
- Bayardo, R. J., & Agrawal, R. (2005). Data privacy through optimal k-anonymization. In *21st international conference on data engineering (icde'05)* (pp. 217–228).
- Chen, Y., & Yang, S. (2007). Estimating disaggregate models using aggregate data through augmentation of individual choice. *Journal of Marketing Research*, *44*(4), 613–621.
- Clementi, F., & Gallegati, M. (2005). Pareto's law of income distribution: Evidence for germany, the united kingdom, and the united states. In *Econophysics of wealth distributions* (pp. 3–14). Springer.
- Gelman, A., & Hill, J. (2006). *Data analysis using regression and multilevel/hierarchical models*. Cambridge university press.
- Granger, C. W. (1981). Some properties of time series data and their use in econometric model specification. *Journal of econometrics*, *16*(1), 121–130.
- Hofmann, D. (1997). An review of the logic and rationale of hlm. *Journal of Management*, *23*(6), 723–742.
- King, G. (1997). *A solution to the ecological inference problem: Reconstructing individual behavior from aggregated data*".
- Koorn, R., Bholasing, J., Pipes, S., Rotman, D., Kypreos, C., Cumming, S., . . . Manchu, T. (2015). *Big data analytics & privacy: How to resolve this paradox*. Compact.
- Lupton, R. (1993). *Statistics in theory and practice*. Princeton University Press.
- McAfee, A., Brynjolfsson, E., Davenport, T. H., Patil, D., & Barton, D. (2012). Big data: the management revolution. *Harvard business review*, *90*(10), 60–68.
- McNeish, D. M. (2014). Analyzing clustered data with ols regression: The effect of a hierarchical data structure. *Multiple Linear Regression Viewpoints*, *40*(1), 11–16.
- Musalem, A., Bradlow, E. T., & Raju, J. S. (2009). Bayesian estimation of random-coefficients choice models using aggregate data. *Journal of Applied Econometrics*, *24*(3), 490–516.
- Narayanan, A., & Shmatikov, V. (2008). Robust de-anonymization of large sparse datasets. In *2008 ieee symposium on security and privacy (sp 2008)* (pp. 111–125).
- Pham-Gia, T., & Hung, T. (2001). The mean and median absolute deviations. *Mathematical and Computer Modelling*, *34*(7-8), 921–936.

- Raghunathan, B. (2013). *The complete book of data anonymization: from planning to implementation*. CRC Press.
- Raudenbush, S. W. (2004). *Hlm 6: Hierarchical linear and nonlinear modeling*. Scientific Software International.
- Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (Vol. 1). sage.
- Rosar, F. (2015). Continuous decisions by a committee: median versus average mechanisms. *Journal of Economic Theory*, *159*, 15–65.
- Ruff, R. R., Akhund, A., & Adjoian, T. (2016). Small convenience stores and the local food environment: an analysis of resident shopping behavior using multilevel modeling. *American Journal of Health Promotion*, *30*(3), 172–180.
- Sapsford, R., & Jupp, V. (1996). *Data collection and analysis*. Sage.
- Scheffer, J. E., & David, E. L. (1985). Estimating residential water demand under multi-part tariffs using aggregate data. *Land Economics*, *61*(3), 272–280.
- Smit, E. G., Van Noort, G., & Voorveld, H. A. (2014). Understanding online behavioural advertising: User knowledge, privacy concerns and online coping behaviour in europe. *Computers in Human Behavior*, *32*, 15–22.
- Steenburgh, T. J., Ainslie, A., & Engebretson, P. H. (2003). Massively categorical variables: Revealing the information in zip codes. *Marketing Science*, *22*(1), 40–57.
- Tong, Y. L. (2012). *The multivariate normal distribution*. Springer Science & Business Media.
- Walker, H. M. (1943). *Elementary statistical methods*.
- Wan, X., Peng, L., & Li, Y. (2015). A review and comparison of methods for recreating individual patient data from published kaplan-meier survival curves for economic evaluations: a simulation study. *PLoS One*, *10*(3), e0121353.
- Wedel, M., & Kamakura, W. A. (2012). *Market segmentation: Conceptual and methodological foundations* (Vol. 8). Springer Science & Business Media.
- Wedel, M., & Kannan, P. (2016). Marketing analytics for data-rich environments. *Journal of Marketing*, *80*(6), 97–121.
- Zenor, M. J., & Srivastava, R. K. (1993). Inferring market structure with aggregate data: A latent segment logit approach. *Journal of Marketing Research*, *30*(3), 369–379.

Appendix

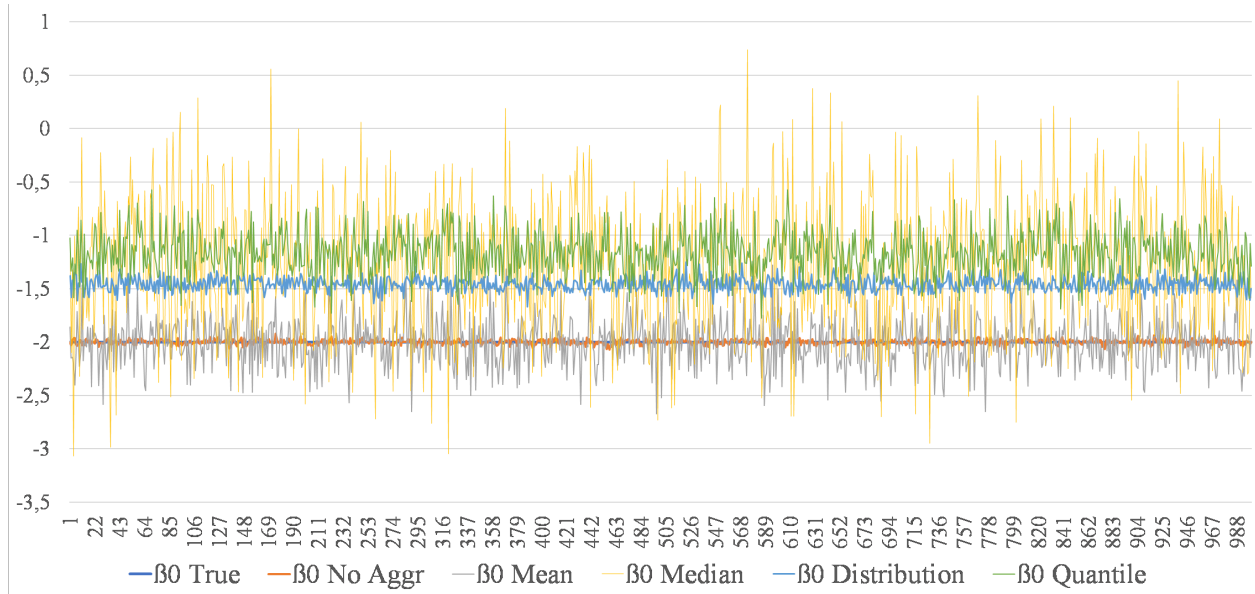


Figure 6: $\beta_0 = -2$ Development over Simulation Runs

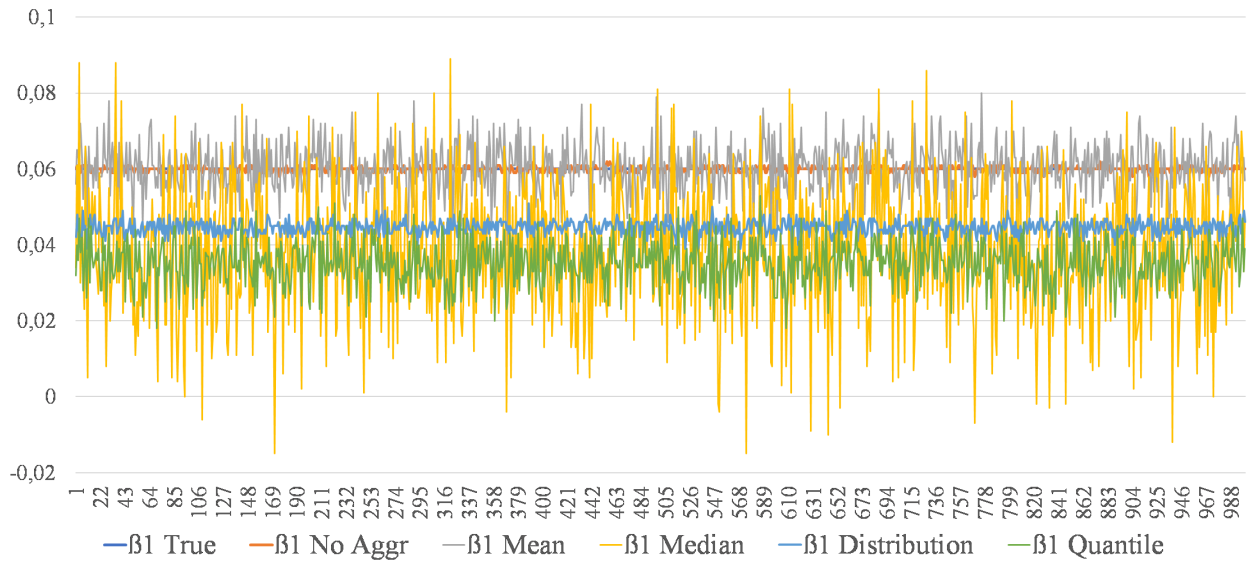


Figure 7: $\beta_0 = 0.06$ Development over Simulation Runs

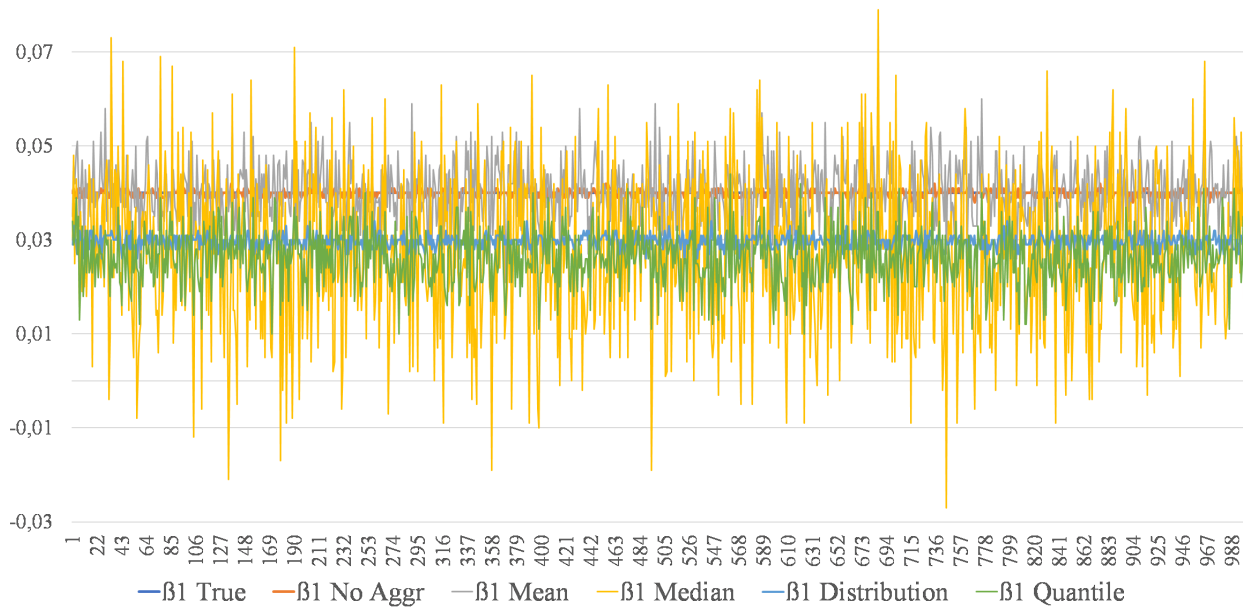


Figure 8: $\beta_0 = -1$ Development over Simulation Runs

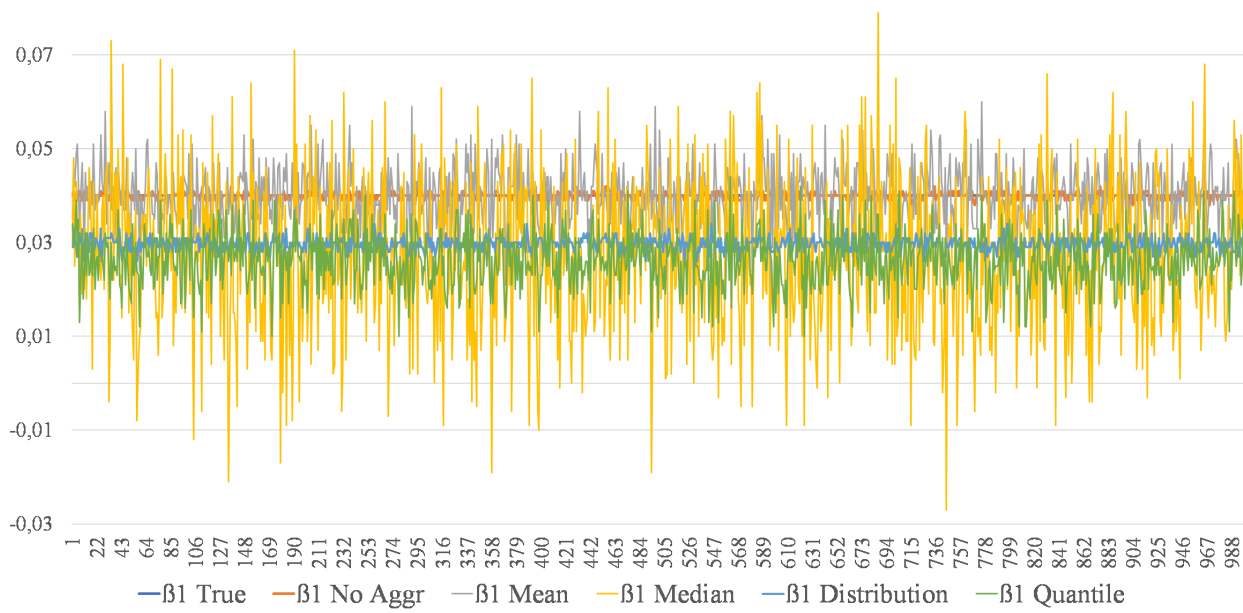


Figure 9: $\beta_1 = 0.04$ Development over Simulation Runs

Table 11: Relative Error Comparison

(lr)	Household Level		Zip Code Level			
	Actual Values	Estimated values	Mean	Median	Distribution	Quantile
β_0						
-1	0.0000%	-0.1184%	-5.2745%	-38.8206%	-58.1179%	-66.1878%
-2	0.0000%	-0.0419%	-0.9896%	-214.9940%	-37.5179%	-78.0800%
-3	0.0000%	-0.0296%	-0.3584 %	219.8380 %	-43.7045%	-74.0098%
β_1						
0.04	0.0000%	-0.0763%	-2.2275%	-66.8372%	-35.0772 %	-62.7922%
0.06	0.0000%	-0.0581%	-0.8752%	-97.3023%	-34.7549%	-76.1017%
0.10	0.0000%	-0.0266%	-0.2133%	-71.4244%	-34.8621%	-72.6447%

Table 12: Hit Rate

	β_0			β_1		
	-1	-2	-3	0.04	0.06	0.1
Mean	86.9%	91.8%	74.3%	86.6%	93.8%	71.8%
Median	11.9%	8.1%	18.1%	10.0%	5.7%	13.0%
Distribution	1.2%	0.0%	7.6%	1.3%	0.0%	5.3%
Quantile	0.0%	0.1%	0.0%	2.1%	0.5%	9.9%