The Impact of Socioeconomic Data on Delivery Time Prediction

Student details

Name: S. van den Boomen
Student number: U428312

THESIS SUBMITTED IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE DEGREE OF
MASTER OF SCIENCE IN DATA SCIENCE & SOCIETY
DEPARTMENT OF COGNITIVE SCIENCE & ARTIFICIAL INTELLIGENCE
SCHOOL OF HUMANITIES AND DIGITAL SCIENCES
TILBURG UNIVERSITY

Thesis committee

Supervisor: dr. A. T. Hendrickson
Second reader: dr. I. Önal

Tilburg University
School of Humanities & Digital Sciences
Department of Cognitive Science & Artificial Intelligence
Tilburg, The Netherlands
09/07/2021

Preface

This thesis was written as the final step to graduation from the Data Science and Society program at Tilburg University. What started as a last-minute change of plans because of a global pandemic has now culminated into this research piece that signals the end of a year of studying from home. A year dictated by online classes, zoom meetings and uncertainty. What illustrates the past year best is that, if this thesis receives a passing grade, I will have graduated from Tilburg University without ever having set foot on the campus itself.

Table of Contents

Abstract

The goal for this thesis was to find how SES data might influence parcel delivery time prediction. A dataset from a bicycle company in Eindhoven (TDV) was used to build a baseline simple regression model to predict parcel delivery time. This dataset was then coupled to a dataset from Statistics Netherlands (CBS) that contained several categories of SES data. From this data regression models were built per SES data category. The algorithms used for these models were Linear Regression and regularized regression models Lasso Regression, Ridge Regression and ElasticNet Regression. This thesis has not found significant results that indicate that SES data has a meaningful contribution to the prediction of parcel delivery time.

1. The impact of socio-economic data on delivery time prediction.

The city of the future is sometimes envisioned as a sustainable green urban oasis. Still, 12% of CO_2 emissions in the Netherlands originate from road transport of which 30 to 35% is linked to urban freight transport (Topsector Logistiek, 2019). Urban freight transport operations such as the stocking of shops, offices and construction sites, the delivery of a parcel containing new clothes and the florist delivering a bouquet of flowers for a birthday thus constitutes to about 4% of CO_2 emissions in the Netherlands. This is far from realizing a sustainable green urban oasis.

The impact of urban freight transport on CO_2 emissions has led to 30 to 40 cities in the Netherlands, including the city of Eindhoven, planning for zero-emission zones (Topsector Logistiek, 2019). Eindhoven has dedicated the area around the center to gradually introduce new emission regulations from 2021. Eventually, the goal is to create a zero-emission zone where only non-emitting vehicles are allowed to enter (Figure 1)[1].

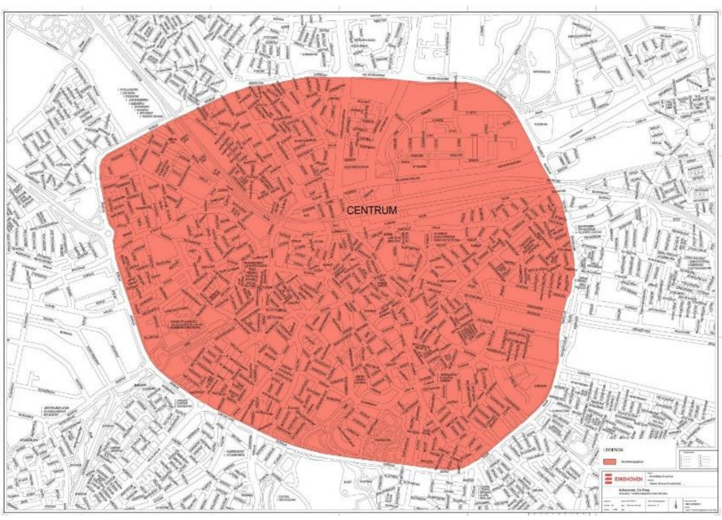

Figure 1 Zero-emission zone in Eindhoven

Making the city center a zero-emission zone reduces local pollution. Though, it could potentially force logistical operations to other parts of the city that in turn experience increased emissions and other downsides such as (noise) pollution. It is therefore important to understand the

[1] Gemeente Eindhoven, Op weg naar een nul-emissiezone. https://www.eindhoven.nl/projecten/nul-emissiezone/op-weg-naar-een-nul-emissiezone

current logistical operations in both the city center and other neighborhoods of Eindhoven, as it should be a goal to make urban freight transport sustainable for the entirety of the city.

Furthermore, it is expected that over 90% of the population in the Netherlands will live in cities by 2050 (Slabinac, 2015). The increase in population will also increase the quantity and diversity of goods that are bought and delivered to customers as cities concentrate population and economic activities (Slabinac, 2015). This means that there is a need for an effective but certainly also sustainable logistical operations policy for the future city.

To understand the needs of the city, it is important to consider that a city in itself is not universal. What works in a certain neighborhood might not work in another. It would not make a lot of sense to have a policy that is designed specifically for an urban area deployed at an industrial site. Though, it is not feasible to create a separate policy for each neighborhood. Therefore, this thesis proposes to use Socio-Economic Status (SES) data to group neighborhoods with similar characteristics. By investigating if SES data and what categories of SES data impact parcel delivery, a better understanding of urban freight transport at a neighborhood level can be achieved. To investigate the impact SES has on urban freight transport, several models will be built, using different SES data categories, with the aim to answer the following questions:

RQ: What can be learned from using SES data for parcel delivery time prediction?

SQ: What features are important when predicting parcel delivery times?

SQ: What is the impact of SES data on parcel delivery time prediction?

> **SSQ:** What is the impact of different categories of SES data on parcel delivery time prediction?

SQ: What model is most useful when measuring the impact of SES data?

To get a representation of current urban freight transport operations, a dataset by Tour de Ville Fietskoeriers, a bicycle messenger company that operates in Eindhoven, is used. This data is linked to a dataset by Statistics Netherlands (CBS) that contains several categories of SES data on a neighborhood level.

For the city of Eindhoven, understanding the dynamics of urban freight transport on a neighborhood level can allow for the finetuning of policies to optimize urban freight transport in the city of the future. As an example, Eindhoven could choose to encourage depots in neighborhoods where delivery takes relatively long. Also, for residents and businesses in a city it is of vital importance to remain accessible, by having an overview of urban freight transport operations at a neighborhood level, the city of the future can be tailored to meet these demands.

2. Related work

Over the past decades, the growth in e-commerce has led to a substantial increase in the number of parcel deliveries (Chu, Zhang, Bai, & Chen, 2021). In Germany, for example, an annual growth rate of 4.7% was expected for parcel deliveries, even before the COVID-19 pandemic was taken into account (Hagen & Scheel-Kopeinig, 2021). Parcel delivery, also known as the last-mile (the shipping of a parcel to its final destination), is considered to be one of the most expensive, inefficient and polluting steps in the supply chain (Gevaers, Van de Voorde, & Vanelslander, 2009) (Wrighton & Reiter, 2016) (Slabinac, 2015) (Archetti & Bertazzi, 2020). Parcel deliveries are an import part of urban freight transport.

2.1 The characteristics of urban freight transport

A common modeling method to measure performance of urban freight transport is known as the Vehicle Routing Problem (VRP). The VRP is a combinatorial optimization problem that tries to find the optimal routes for delivering goods for a given set of delivery vehicles operating from a depot (Golden, Raghavan, & Wasil, 2008). There exist many variants of the VRP (Golden, Raghavan, & Wasil, 2008) as different characteristics associated with the last-mile delivery influence the optimal solution for the VRP. Gevaers, Van de Voorde and Vanelslander (2009) distinguish five characteristics of innovations in last-mile delivery that impact VRP optimization:

1. *The level of service offered to the consumer:* Services that are offered to consumers are, for example, time windows, maximum lead times, the frequency with which parcels are delivered to an address and routes where goods that have to be returned are picked up. Usually, the more services are offered, the less efficient the routing becomes. The study shows, for example, that smaller time windows increase distance that needs to be travelled per stop which in turn increases pollution.
2. *The security and delivery type:* Some deliveries have to be attended or require a signature from the recipients while others can be put in the mailbox or are sent to a collection point or delivery box. This impacts the efficiency and optimization of a delivery route.
3. *The Geographical area & Market penetration:* A delivery route in an area with a high population density and large market share for the company performing the deliveries are more efficient when it comes to route optimization. Geographical features and market share can thus have an impact on routing efficiency.

4. *Fleet & Technology:* The effectiveness of the delivery fleet (fuel load, capacity, loading type etc.) and the IT-systems used for delivery can have a major impact on route optimization as efficient fleets and systems improve performance.
5. *The Environment:* The environment also plays a role in route optimization as, for example, some of the above mentioned aspects might have negative impact on pollution (e.g. smaller time windows causing longer travel distance between stops which results in more air-pollution).

While the research goal for this thesis is not to come up with a solution in the form of a VRP model, the categories do give an indication of what can have an influence on the prediction of parcel delivery time. For example, Gevaers, Van de Voorde & Vanelslander (2009) state that the population density of a neighborhood could impact the efficiency of parcel delivery.

A paper by Cruz de Araujo and Etemad (2021) uses a dataset from Canada Post that contains 6 months of deliveries performed in the Greater Toronto Area, Canada, to predict parcel delivery time. With the inclusion of GPS and weather data, they train several Deep Learning models (Cruz de Araujo & Etemad, 2021). They show that their Deep Learning models perform better than normal Machine Learning models but, because of their black-box nature, they can only look at the errors to see how the model comes to a prediction (Cruz de Araujo & Etemad, 2021). This would make causal analysis using the features less effective. Something that is useful however, is that they use Mean Absolute Percentage Error (MAPE) as a metric (Cruz de Araujo & Etemad, 2021), which would be a good option to use here as well.

A paper that uses Machine Learning methodologies to predict stop delivery times is that of Hughes, Moreno, Yushimito and Huerta-Cánepa (2019). Their paper compares both regression and classification algorithms to predict if stop times would exceed a certain threshold, though, as they are more interested in the time spend at the actual stop rather than the time travelled it is not very relevant for this paper, however the inclusion of MAPE as a metric is another incentive to use it for model comparison here as well (Hughes, Moreno, Yushimito, & Huerta-Cánepa, 2019).

2.2 The negative impacts of urban freight transport

As mentioned before, 12% of CO_2 emissions in the Netherlands originate from transport of which 30 to 35% can be linked to urban freight transport (Topsector Logistiek, 2019). Urban freight transport is thus negatively impacting the environment, which also shows in the labelling of *The Environment* as a characteristic by Gevaers, Van de Voorde & Vanelslander (2009). Though, this is still a rather broad category. Slabinac (2015) further dissects the negative impacts generating from urban freight transport into four categories:

1. *Negative environmental impacts:* The depletion of non-renewable resources, air pollution as well as various other sources of waste such as used up tires, vehicles, and unsustainable packaging material among others.
2. *Negative social impacts:* Aspects of urban freight transport that negatively impact Quality of Life. This includes negative impacts on public health such as deaths or injuries sustained from traffic accidents, nuisances that arise from pollution (e.g. air, noise, vibration or visual pollution) and physical threats and intimidation by the size of the transport vehicles.
3. *Negative economic impacts:* Impacts associated with road congestion and economic burdens to stakeholders involved in urban freight transport because of inefficiencies and the negative environmental and social impact that urban freight transport has (Slabinac, 2015). The cost of traffic congestion alone, constitutes to nearly 100 billion Euros annually

in Europe, which is about 1% of the GDP of the European economy (Bektas, Crainic, & van Woensel, 2015).

4. *Negative operational impacts:* Negative operational impacts refer to congestion and traffic disruptions. The (un)loading, parking and maneuvering of vehicles, as part of urban freight transport, can block or hinder other users or delivery services which in turn can have negative impact on operating results.

The WHO (World Health Organization, 2005) also recognizes the impact of air-pollution generated by transport on health outcomes. The impacts on health include mortality, non-allergic respiratory morbidity, allergic illness and symptoms of allergic illnesses (e.g. asthma), cardiovascular morbidity (e.g. heart attacks), cancer, pregnancy and birth outcomes (e.g. premature birth and miscarriages) and male fertility (World Health Organization, 2005). Reducing air-pollution have shown to directly reduce acute asthma attacks in children. In the long term, life expectancy is expected to rise and the annual number of deaths contributed to respiratory and cardiovascular disease are expected to lower (World Health Organization, 2005).

Therefore, there are multiple incentives for improving urban freight transport. Research can be done with economic gain as a goal by increasing efficiency, as can the focus be on the reduction of pollution and the prevention of other negative impacts. Usually, research is conducted aiming for a combination of both economic gain and negative impact reduction. This also increases the validity of the solutions in terms of real-world applicability.

2.3 Innovations in urban freight transport

The negative impacts associated with urban freight transport have led to a push for innovation in the logistics field. An example of an innovative approach is a study by Ohsugi & Koshizuka (2018) where the real-time energy usage of households was used to build a model that could predict if there was someone at home, as recipient absence has a major impact on efficiency and pollution. This study showed promising results with an 87.5% reduction of absent package deliveries (Ohsugi & Koshizuka, 2018). Information on energy usage could thus be an impactful predictor for estimating parcel delivery time.

Innovations can also be a direct response to legislative action that aims to combat the negative impacts. As an example, cargo bicycles can be used to circumvent restrictive access protocols for motorized vehicles (Naumov, Vasiutina, & Solarz, 2021). Restrictive access policies, such as the proposed plan to close of the city center in Eindhoven for vehicles other than emission-free ones, also makes way for new ways of distribution. One option is to, instead of having vans travel from depots outside of the city to their delivery areas, have satellite depots that can be set-up at strategic locations in urban areas from which deliveries can be performed.

There are two types of these satellite depots, Urban Freight Mini-hubs and Urban Micro-consolidating Centers (UMCs) (Muñuzuri, Cortés, Grosso, & Guadix, 2012). The main difference is that Urban Freight Mini-hubs consists of designated areas, such as parking spaces, where a distribution van can park regardless of access times and can deliver goods on foot or with a handcart, whereas UMCs are small urban distribution centers where larger quantities of goods can be brought for further distribution by small EV's or cargo bikes (Muñuzuri, Cortés, Grosso, & Guadix, 2012). A UMC is typically located close to, or in the urban area (Muñuzuri, Cortés, Grosso, & Guadix, 2012). Another paper looked at UMCs that are located in Paris and London in order to see if it was a feasible option for Manhattan (New York) which suffered from severe urban transport challenges due to congestion and overcrowding (Conway, Fatisson, Eickemeyer, Cheng, & Peters, 2012). A proposed solution was to use cargo tricycles to deliver parcels from a UMC. Also, UMCs should be accessible for multiple companies that could then leave the parcel delivery to the tricycles at the location instead of having their vans go into the city themselves. The feasibility of the location of the UMCs was investigated using information on bus lane miles, bicycle lane miles, building, office, industry, and retail space as well as the assessed value of the building spaces (Conway, Fatisson, Eickemeyer, Cheng, & Peters, 2012). It could be interesting to see if SES data on buildings and industry impacts prediction, with potential for translating this information into guidelines for new UMCs.

The use of cargo bicycles or cargo tricycles in urban freight transport has also been the subject of research over the past decade. The *Cyclelogistics* and *Cyclelogistics Ahead* projects in Europe are examples of the successful implementation of cycling based urban freight transport (Wrighton & Reiter, 2016). The projects showed that last-mile delivery by cargo bike (€1.60 per parcel) is more profitable in densely populated areas than conventional delivery with motorized vehicles (€2.91 per parcel), in addition to being better in terms of environmental impact (Wrighton & Reiter, 2016).

2.4 The significance of Socio-economic Status in prediction

So far, there are some indications that data that relates to Socio-economic Status (SES) can be used for urban freight transport modelling. For example, the population density might influence delivery efficiency (Gevaers, Van de Voorde, & Vanelslander, 2009), energy usage monitoring can drastically reduce recipient absence (Ohsugi & Koshizuka, 2018) and the number of offices or retail locations might have an impact on the feasibility of a UMC location (Conway, Fatisson, Eickemeyer, Cheng, & Peters, 2012).

SES has been a valuable scientific data source. Numerous studies have been performed using SES data. An extensive literature review on obesity and SES by McLaren (2007), for example, analyzed 333 studies linked to SES data published between 1988-2004. Another example is a study that investigated the relationship between SES data (income level, employment status, environmental status and educational attainment), and cardiovascular disease (Schultz, et al., 2018). While this is not directly relatable to this research project, it does show that SES can be a valuable data source.

An example of a Machine Learning (ML) approach to using SES data, is a paper predicting a women's height from their respective SES (Daoud, Kim, & Subramanian, 2019). The paper compared the performance of seven ML methods (Lasso regression, RIDGE regression, generalized additive model, Bayesian Neural Net, bagged CART, and Random Forest) to OLS regression. The paper concluded that, while Bayesian Neural Net performed best in terms of explained variance, this improvement was only marginal (0.3%) in comparison with OLS regression. Daoud, Kim & Subramanian (2019) saw this as an indication that there were no non-linear relationships between SES and height. Furthermore, the paper recommends reporting the feature significance in prediction, as models that are transparent when it comes to the impact of features, offer more insights when performing causal analysis (Daoud, Kim, & Subramanian, 2019).

As it is a goal of this thesis to find the impact of (categories of) SES data on prediction, it is important to consider the feedback a model provides in terms of causal analysis. While a Deep Learning model might perform better in terms of prediction accuracy, it's black box nature might reduce the applicability when it comes to causal analysis compared to, for example, a regression analysis.

Overall, there are some indications that SES data can have an impact on prediction. Though, there has not yet been a research project that specifically looks for these relations or at the different kinds of SES data available to make predictions and evaluate them in terms of causal analysis. This paper hopes to contribute to innovations in urban freight transport by investigating the impact SES data has, potentially opening up new avenues of research for future projects.

3. Methodology

It is not the aim to classify parcel deliveries in subgroups, rather it is the goal to predict parcel delivery time. Therefore, regression algorithms will be used instead of classifying algorithms. The first series of models will be based on linear regression. For the baseline, a simple linear regression model is used, parameters are estimated using OLS estimation.

In Machine Learning (ML), the goal is to find a model that not only predicts well on the training data, but also performs well on similar data that was not used to train the model. The introduction of a small amount of bias to a model can improve variance and the performance of the model on non-training data (Daoud, Kim, & Subramanian, 2019). This is also known as the bias-variance trade-off. In ML regularization techniques are used to regulate the bias-variance trade-off.

3.1 Regularization

Because the SES data introduces several variables, there is a risk of overfitting, especially as there is considerable multicollinearity between features. Linear regression has the tendency to pick up on trends that are only present in the training data as its goal is to minimize bias in this training set. With the introduction of regularization methods, a small amount of bias is added which should reduce the variance, and thus overfitting, of the model. For this thesis three regularization techniques are considered: Least Absolute Shrinkage and Selection Operator (Lasso), Ridge and Elastic Net regularization. The impact of the regularization term is controlled by setting the value for λ. λ can range from 0 to +∞. When the value of λ is set to 0, the original parameters observed from OLS estimation are obtained. The larger the value for λ, the more the parameters of the model are penalized. The exact penalty depends on the regularization technique that is used.

3.1.1 Lasso regression

Lasso regression uses L_1 regularization that introduces an error term (*Figure 2)*, the sum of the absolute coefficients, to the OLS estimation. The size of the error is determined by the hyperparameter λ, which can be tuned to obtain optimal performance, *β* represents the coefficients. Lasso regression pushes coefficient estimates that have a smaller contribution to the model to zero. Lasso tends to perform best on data where there are some predictors with large coefficients and some smaller, less important, ones. These coefficients are then pushed to zero, which is also a kind of feature selection as predictors that are important in prediction remain and predictors that are not important are reduced to zero. A downside of Lasso could be that, when dealing with correlated features, it tends to favor one feature over the others.

$$+\lambda \sum_{j1}^{k} |\beta_j|$$

Figure 2 the L_1 regularization term added to the OLS estimation.

3.1.2 Ridge regression

Ridge regression uses L_2 regularization that introduces an error term (*Figure 3)*, the sum of the squared coefficients, to the OLS estimation. The size of the error is determined by the hyperparameter λ, which can be tuned to obtain optimal performance, *β* represents the coefficients. In contrast with Lasso (L_1 regularization), in Ridge regularization (L_2 regularization) the value of the coefficients cannot become 0, though it can be pushed close to zero. Ridge thus cannot completely remove features from a model, though it can reduce their influence. This does mean that it is not as useful in terms of feature selection and might prove less impactful for the research goals.

$$+ \lambda \sum_{j}^{k} \beta_j^2$$

Figure 3 the L_2 regularization term added to the OLS estimation

3.1.3 Elastic Net

Elastic Net regularization combines L_1 and L_2 regularization which is again controlled by setting a value for λ, which can be tuned to obtain optimal performance. Next to λ, the ratio at which both regularization techniques are used can be set and tuned. By combining both regularization types, it effectively shrinks coefficients (L_2) as well as setting some to zero (L_1). This might be useful as some of the features that get lost in Lasso remain, while it still pushes other coefficients to zero.

4. Experimental Setup

4.1 The Tour de Ville Eindhoven dataset

For this thesis, a dataset from Tour de Ville Fietskoeriers Eindhoven (TDV) is used that contains an overview of urban freight transport operations in Eindhoven performed by TDV. TDV is a bicycle messenger company operating from Eindhoven that performs logistical operations by bike, providing both business to business (B2B) and business to consumer (B2C) logistical services in Eindhoven and the surrounding area.

The TDV dataset is not publicly available, and access was granted exclusively for this thesis. The raw dataset contains 22482 rows with 42 features that contains all logged orders performed by TDV between 01/11/2020 and 24/03/2021. An example highlighting the most important features can be found in *Appendix A*.

During initial cleaning, features that had privacy sensitive data (e.g. e-mail addresses) or contained redundant information (e.g. comments added by messenger on pick-up or delivery) were dropped. Also instances that were completed by fictive employees for administration purposes (e.g. completed by a messenger named ‘Admin’) or had no address information (i.e. no pick-up and delivery address) were dropped. The resulting dataset has 25 features, that are related to address information, messenger and status, for a total of 20005 instances that are used for further processing.

This rough cleaning did not consider validity of the data. It is expected that several instances are invalid. This is in the nature of the order planning and completion process. For order completion, a messenger completes the order in an application called Veloyd. While it is encouraged to do this at the delivery or pickup location, this is not always done correctly. For example, Messenger A returns after completing his route and then signs off all his deliveries instead of at their respective stops. This can lead to invalid values for either *Delivery_at* or *Pickup_at*.

Furthermore, not all stops on a route are logged, especially in the morning and the afternoon where a messenger can be asked to deliver or pick-up an order during a mail delivery or mail retrieval route. The stops on these mail-routes are not logged in Veloyd, which inflates time travelled between the stops that are logged. For example, Messenger A has a mail route with 5 stops and is asked to deliver two packages after his first and fourth stop, Veloyd would only log the data for these two package delivery stops and not consider extra time and distance because of the stops in the mail route. A method to filter out these invalid instances will be discussed later in this chapter.

4.1.1 Extracting the *Traveltime*

Traveltime is a feature that contains the time between the completion of two orders. To extract *Traveltime* from the TDV dataset, first a *Ride_ID* is assigned for each unique combination of *Date* and *Messenger*. Next, the *Delivery* feature is created that sets the timestamp at which the order was completed. By default, the timestamp of the feature *Delivery_at* is taken, unless it is missing, then the timestamp of the feature *Pickup_at* is used. An exception is made for instances for which the delivery address is that of TDV and *Pickup_at* is not missing. In this case, the timestamp for *Delivery* is set as the timestamp for the *Pickup_at* feature. This is done because these instances are pickup orders and not delivery orders which means that the *Pickup_at* feature contains the correct time a messenger completed this order in their route.

Each ride is then chronologically sorted using the *Delivery* timestamp, the *Stop* feature is numbered accordingly. From this *Traveltime* can be calculated by determining the difference in time between two stops (i.e. *Traveltime* for *Stop* 2 is the time difference in minutes between *Stop* 1 and *Stop* 2). For the first *Stop* of a ride, the time difference between *Pickup_at* and *Delivery_at* is taken because no previous timestamp is available to calculate the difference.

4.1.2 Calculating Distance

The baseline model uses *Distance* as the independent variable. To calculate *Distance*, first the house number, address, city, and postal code are extracted. Again, by default the delivery address is taken unless it is not available, then the pickup address is used, as it was a pick-up order without delivery. As for the *Delivery* timestamp, the same exception applies that, if delivery address is that of TDV, the pickup address is used instead.

4.1.2.1 Bing Maps REST Service

To calculate the distance between two addresses, the Bing Maps Routes API[2] is used for which a key was obtained via an educational license. A GET request was sent to obtain the walking distance between two addresses which was subsequently stored as *Distance*. The routing API does not have a cycling option. Therefore, the walking option is used as it most closely represents the cycling distance as cars often have one-way roads or roads that restricted for foot and bike traffic.

4.2 The CBS Socio-economic Status dataset

The TDV dataset is linked to a dataset from Statistics Netherlands (CBS) that contains Socio-Economic Status (SES) data at neighborhood level. The source for the SES data is the 'Kerncijfers wijken en buurten 2019' dataset from Statistics Netherlands (CBS).[3] The features in this dataset can be classified in the following SES categories: Population, Living, Energy, Education, Labor,

---

[2] Documentation available at: https://docs.microsoft.com/en-us/bingmaps/rest-services/routes/calculate-a-route
[3] Available at: https://www.cbs.nl/nl-nl/maatwerk/2019/31/kerncijfers-wijken-en-buurten-2019

Social Security, Care, Business Locations, Motor Vehicles, Services, Surface, Postal Code and Urbanity. Each category has multiple features.

The 'Kerncijfers wijken en buurten 2019' dataset does not contain address information but uses a *Buurtcode* (neighborhood code) instead. In order to obtain the *Buurtcode* for a given address, the 'Buurt, wijk en gemeente 2020 voor postcode huisnummer' dataset, also available from CBS, is used.[4] From this dataset *Buurtcode* for each instance was extracted using postal code and house number.

The obtained *Buurtcode* is then used to extract the SES data for an instance from the 'Kerncijfers wijken en buurten 2019' dataset. Some of these features are in absolute numbers, which makes them incomparable between neighborhoods. This is solved by transforming them to percentages by using the number of inhabitants of a neighborhood. A description of the features in the final dataset can be found in *Appendix B*.

4.3 Data selection boundaries

With the *Traveltime* and *Distance* features, average speed can be calculated (*KM/h)*. By selecting the instances for which *KM/h* is between 10 km/h and 30 km/h, some of the invalid instances are excluded from the data. These boundaries are set because it is highly unlikely that a messenger cycles faster than 30 km/h including time necessary to deliver or pickup an order. The lower limit at 10 km/h should filter out some of the stops that have *Traveltime* inflated by unlogged or incorrectly completed orders.

[4] Available at: https://www.cbs.nl/nl-nl/maatwerk/2020/39/buurt-wijk-en-gemeente-2020-voor-postcode-huisnummer

Some of the orders that are outside of the Eindhoven area have been completed using other means of transport and fall within the boundary set for *KM/h*. To filter them out, instances for which *Distance* is over 18 kilometer and/or are not completed in Eindhoven or surrounding villages (Veldhoven, Best, Waalre, Son en Breugel, Nuenen, Son, Mierlo or Geldrop) are dropped from the data as well. After this, a dataset containing 4055 instances remains.

4.3.1 Neighborhood imbalance

In A*ppendix C*, the neighborhood frequency distribution can be found. The distribution shows a difference between residential and industrial neighborhoods. The neighborhood that is most frequent in the dataset, 'Hurk', only has 70 inhabitants, and the 7^{th} most frequent neighborhood 'Flight Forum' has no inhabitants. This also means that these cases can have inflated or missing SES data. For example, the *Percent_youth_services* are 40% for the 'Hurk' neighborhood which is high compared to the mean (μ = 0.097). Industrial neighborhoods are therefore filtered by introducing the Boolean *Industrial* feature. This feature uses the value for *Businesses_per_inhabitant* to classify a neighborhood as industrial. The cut-off point is set at equal or more than 0.52 businesses per inhabitant. This was done to classify the 'Strijp-S' neighborhood, a neighborhood that houses a mix of business, retail and housing, as non-industrial (*Businesses_per_inhabitant['Strijp-S'] = 0.51)*. Also, neighborhoods that have no inhabitants are classified as *Industrial*. In total, 416 instances are classified as *Industrial*. When verifying with the *Percent_youth_services* feature, the mean has dropped significantly (μ = 0.076 vs μ = 0.097, for the dataset excluding *Industrial* neighborhoods). The exclusion of instances in *Industrial* neighborhoods also reduces the number of missing values as they most often occur for neighborhoods classified as *Industrial*.

4.4 Missing data treatment

Eliminating neighborhoods classified as *Industrial* does not remove all missing data. To treat missing data for neighborhoods, multiple imputation is used. First a dataset is created that contains all features that have no missing instances. Next, features that have incomplete data and for which the number of incomplete instances is less than 350 are listed. Features with over 350 missing instances are regarded as unfit as too much data would have to be imputed, these features are: *Avg_energy_usage_semidetached, Avg_energy_usage_detached, Avg_gas_usage_apps, Avg_gas_usage_semidetached, Avg_gas_usage_detached* and *Percent_district_heating.*

The estimation of the missing features is done sequentially by adding a single feature with missing data to the complete dataset. The *IterativeImputer*[5] is then trained on the subset of the complete dataset that has complete data for the feature that has been added with missing instances. The trained *IterativeImputer* is then used to estimate the missing instances in the complete dataset. This process then repeats until all the features with missing data are treated. The estimation is done with a *BayesianRidge* algorithm based on a round-robin process with a max number of 100 iterations.

---

[5] Documentation can be found at: https://scikit-learn.org/stable/modules/generated/sklearn.impute.IterativeImputer.html#sklearn.impute.IterativeImputer

4.5 Data transformation

The density plot for the *Distance* feature (*Figure 4*) shows considerable skew to the right as well

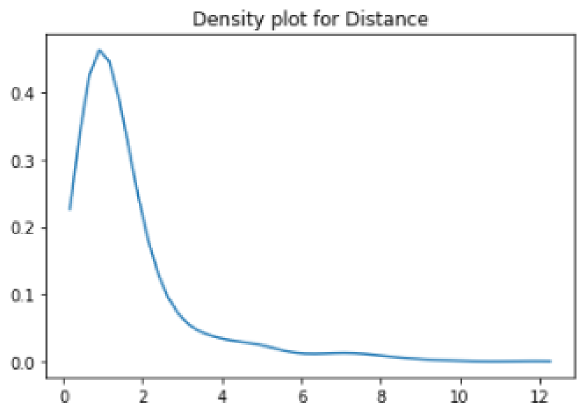

Figure 4 Density plot for the *Distance* feature.

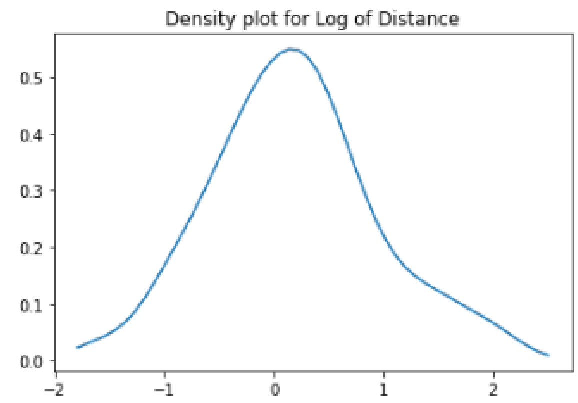

Figure 5 Density plot for the *Log_Distance* feature.

as being leptokurtic. Testing for skewness (2.559)[6] and kurtosis (7.806)[7] confirms this. A solution could be to implement a logarithmic transformation of the *Distance* feature (*Figure 5*). The *Log_Distance* feature does show better skewness (0.264) and kurtosis (0.071) which can be considered as decent. Though, when testing both for normality using the Shapiro-Wilk test normality is rejected (*Distance*: $W(3634) = 0.717$, $p < 0.000$ and *Log_Distance:* $W(3634) = 0.991$, $p < 0.000$). Also, performing logarithmic transformations on the *Distance* feature might interact

---

[6] Documentation available at: https://pandas.pydata.org/pandas-docs/stable/reference/api/pandas.DataFrame.skew.html

[7] Documentation available at: https://pandas.pydata.org/pandas-docs/stable/reference/api/pandas.DataFrame.kurtosis.html

with the linearity and the performance of the linear regression models. This is something to consider when reviewing the model performance.

4.6 Data normalization

Because the features have different scales, for example *Distance* is in Kilometers, and *Percent_men* is a proportion, data normalization is applied. The features are normalized using min max normalization where they are scaled in proportion to the minimum and maximum value a given feature can have.

4.7 Evaluation methods

The first step is to split the data into a training and test set. For this the *Train_test_split* is used,[8] creating a 70/30 train-test split. The hyperparameter tuning is done with *GridSearchCV* to find the *alpha* for λ, and the optimal proportion of L_1 regularization.[9] The cross-validation method used in *GridSearchCV* is *RepeatedKFold.* with 10 splits, 3 repeats and *RMSE* as performance indicator.[10] This results in a set of optimal hyperparameters, λ for Lasso, Ridge and ElasticNet regression, and the proportion of the L_1 regularization for ElasticNet regression.

After hyperparameter tuning, the model is fit on the training data with the optimal hyperparameter settings. This model is then tested on the testing data. Performance is measured using the *RMSE*, R^2 and *MAPE* metrics.

[8] Documentation available at: https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.train_test_split.html
[9] Documentation available at: https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.GridSearchCV.html
[10] Documentation available at: https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.RepeatedKFold.html

5. Results

This section will discuss the performance of the models that were built to assess SES data. First, a baseline model, that excludes SES data, is used to set the base performance indicators. After, models are built per category of SES data and its performance is compared between categories as well as to the baseline model. Comparison between models is done by looking at R^2, *RMSE* and *MAPE* as well as the hyperparameter settings. For causal analysis purposes, feature coefficients are also listed.

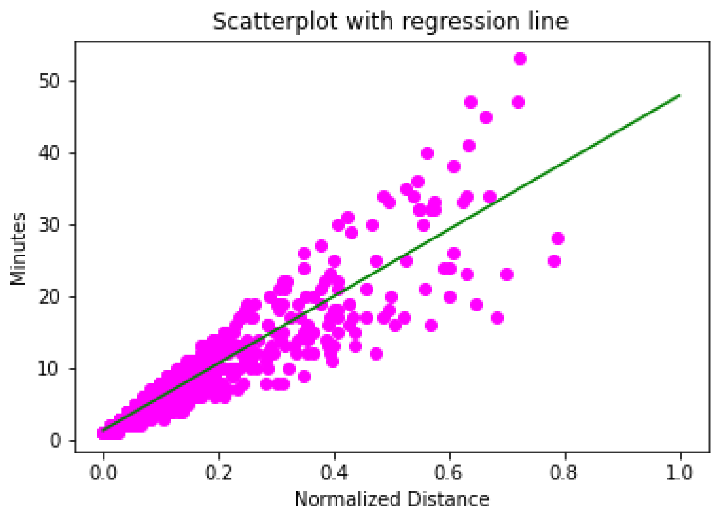

Figure 7 Scatterplot of deliveries including the baseline model regression line.

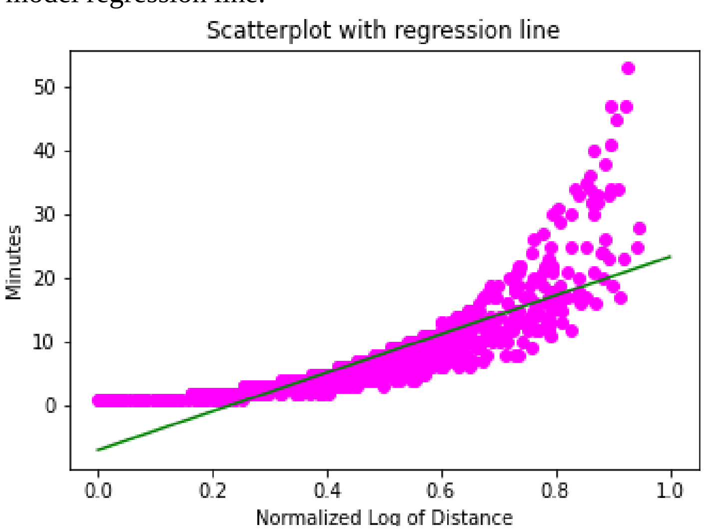

Figure 6 Scatterplot of deliveries including the *Log_Distance* model regression line.

5.1 The baseline model

The baseline model is a simple regression model that uses *Distance* as a predictor for *Traveltime*. The baseline model has a *RMSE* of 2.1686, a R^2 value of 0.8844 and a *MAPE* of 17.082. *Figure 7* shows a scatterplot with regression line for the baseline model. The scatterplot shows considerable skewness and kurtosis as discussed before. This could also lead to the relatively high R^2. Therefore, a log transformation of the *Distance* feature has been performed. This results in the scatterplot that can be seen in *Figure 6*.

From this, it becomes clear that, while there is a significant reduction in skewness and kurtosis, as discussed in the previous section, the linearity of the *Distance* feature is lost. This means that coefficients obtained from training models using *Log_Distance* instead of *Distance* are unfit for performing causal analysis, as when they are transformed back, their meaning is lost. Also, the regression line would lead to predictions that are below 0, which is impossible when it comes to ecological validity. This means that, for training the *Linear Regression, Lasso, Ridge* and *ElasticNet* models it is, in terms of causal analysis, better to use the *Distance* feature.

5.2 Models with SES data

In the following section, the models created per category of SES data will be discussed. This should result in an overview if and/or what categories are useful in prediction and what features might impact delivery time prediction most.

5.2.1. Population model

Table 1 shows the performance and hyperparameters of the models based on Population SES data in comparison to the baseline model. For this model, the *Ridge* regression model performs best in terms of *RMSE* and R^2. In terms of *MAPE* the baseline model performs best, though it has to be said that the models are optimized on *RMSE* and not *MAPE*. *Lasso* and *ElasticNet* models perform significantly worse in terms of *MAPE*. This could be caused by inherent problems that are attributed to *MAPE* as a metric (Davydenko & Fildes, 2016). Furthermore, the proportion of L_1 is high, which indicates that the *ElasticNet* model performs best when using L_1 regularization over L_2 regularization which is an indication that removing features from the model does not harm performance in terms of *RMSE*, indicating that using just the *Distance* feature as a predictor already leads to high performance.

Table 1: **Model performance and hyperparameters for Population SES data.**

Model	RMSE	R^2	MAPE	Alpha	L_1 Proportion
Baseline	2.1686	0.8844	17.082	-	-
Linear Regression	2.1670	0.8846	17.295	-	-
Lasso Regression	2.1653	0.8848	102.173	0.005	-
Ridge Regression	2.1618	0.8851	17.293	0.423	-
Elastic Net	2.1646	0.8848	101.676	0.005	0.990

Table 2 shows the feature coefficients per model. The *Percent_women* feature is excluded as it is 1 – *Percent_men*. When looking at the *Lasso* and *ElasticNet* models, a lot of coefficients have been pushed to zero, again indicating that the *Distance* feature is highly important. This also raises the question to what extent the rest of the coefficients are reliable. Also comparing the size of the coefficient for *Distance* to those in the *Ridge* model shows how influential the *Distance* feature is.

When considering the *Lasso* and *ElasticNet* models, *Percent_migration_western, Percent_migration_non_western(Turkije), Births(per_1000), Percent_1person_hh* and *Population_density_sqkm* are most influential. The effects are similar for the *Linear* and *Ridge* models.

The coefficients for the Linear and Ridge regression models in *Table 2* suggest the presence of some trends. For example, when a neighborhood has a higher percentage of elderly (65+ years old), the expected delivery time is lower, possibly because they are often no longer employed and therefore home more often leading to less time spent per stop. Also, the coefficients on *Moroccan, Antilles* and *Surinam* migratory background show a decrease in delivery time for a higher population with this migratory background. A theory could be that labor participation for these migrant groups is lower (Centraal Bureau voor de Statistiek, 2020) and therefore they can be expected to be home more often, however, the *Percent_migration_non_western* feature, which takes into account other ethnicities as well, and the group with a *Turkish* background do not show this effect.

Table 2: **Feature coefficients for models trained on Population SES data.**

Feature	Linear	Lasso	Ridge	ElasticNet
Distance	46.8747	46.5573	46.3917	46.4124
Percent_men	1.1449	-	0.5812	-
Percent_0_15age	0.3120	-	0.2761	-
Percent_15_25age	1.1974	-	0.8927	-
Percent_25_45age	0.6571	-	0.4535	-
Percent_65+age	-0.7811	-	-0.8736	-
Percent_not_married	-26.8416	-	-1.0477	-
Percent_married	-22.3675	-	-0.4536	-
Percent_divorced	-4.7041	-	0.6642	-
Percent_widowed	-16.0765	-	-0.4170	-
Percent_migration_western	-1.3273	-0.2606	-1.1010	-0.2700
Percent_migration_non_western	0.3032	-	0.1752	-
Percent_migration_non_western(Morocco)	-0.3020	-	-0.2498	-
Percent_migration_non_western(Antilles)	-0.0862	-	-0.0800	-
Percent_migration_non_western(Suriname)	-0.5874	-	-0.3715	-
Percent_migration_non_western(Turkije)	0.2904	0.1777	0.2735	0.1672
Births(per_1000)	-0.6961	-0.4935	-0.9114	-0.4975
Deaths(per_1000)	0.5829	-	0.7897	-
Percent_1person_hh	-0.9625	-0.0238	-0.2287	-0.0243
Percent_no_kids_hh	-0.0642	-	-0.0397	-
Percent_w_kids_hh	0.1579	-	0.1989	-
Avg_size_hh	-0.0586	-	-0.0913	-
Population_density_sqkm	-0.4975	-0.2354	-0.3734	-0.2291

Because *Distance* is such a good predictor, it is possible to look at the residuals from the baseline model and see if the SES data features can predict these residuals well. *Table 3* illustrates the R^2 values obtained from these models. It does confirm that *Distance* is a very good predictor, and that the Population SES data has a negligible contribution to performance.

Table 3 **R^2 values for models predicting residuals of the baseline model on population SES data.**

	Linear	**Lasso**	**Ridge**	**ElasticNet**
R^2	0.0015	-0.0001	0.0037	-0.0001

5.2.2 Living model

Table 4 shows the performance and hyperparameters of the models based on Living SES data in comparison to the baseline model. In similar fashion to the Population model, the performance gain from the inclusion of Living SES data is negligible. *Lasso* and *ElasticNet* models show a very slight improvement in terms of *RMSE* and R^2 but perform much worse in terms of *MAPE*.

Table 4 **Model performance and hyperparameters for Living SES data.**

Model	**RMSE**	**R^2**	**MAPE**	**Alpha**	**L_1 Proportion**
Baseline	2.1686	0.8844	17.082	-	-
Linear Regression	2.1776	0.8834	17.642	-	-
Lasso Regression	2.1666	0.8846	101.98	0.005	-
Ridge Regression	2.1763	0.8836	17.549	0.005	-
Elastic Net	2.1668	0.8846	101.86	0.004	0.99

The coefficients in *Table 5* confirm that the *Distance* feature is the most influential. In the *Linear* and *Ridge* models, the coefficients for *Percent_owner_inhabited,* *Percent_housing_corporation_rental_properties, Percent_rental_properties_other_owners* also show a large effect compared to the rest of the features.

Table 5 **Feature coefficients for models trained on Living SES data.**

Feature	Linear	Lasso	Ridge	ElasticNet
Distance	46.8256	46.5385	46.8264	46.4857
Housing_stock_per_inhabitant	-0.6209	-0.2443	-0.6196	-0.3231
Avg_price_home(x1000)	-0.1640	-	-0.1611	-
Percent_1family_housing	-0.2454	-	-0.2047	-
Percent_inhabited	0.7526	-	0.7824	0.0376
Percent_owner_inhabited	-21.4025	-	-14.9387	-
Percent_housing_corporation_rental_properties	-27.2827	-0.0691	-19.1363	-0.0934
Percent_rental_properties_other_owners	-29.0218	0.4048	-19.8714	0.5284
Percent_owner_unknown	-0.8716	-0.2277	-0.7193	-0.2788
Percent_homes_build_before_2000	0.2399	-	0.2092	-

When using the Living SES data features to predict on the residuals from the baseline model (*Table 6*) it shows that performance is worse than plotting a horizontal line as the R^2 values are negative, the Living SES data is therefore unable to explain the variance in the residuals.

Table 6 **R^2 values for models predicting residuals of the baseline model on living SES data.**

	Linear	Lasso	Ridge	ElasticNet
R^2	-0.0086	-0.0001	-0.0075	-0.0001

5.2.3 Energy model

Table 7 shows the performance and hyperparameters of the models based on Energy SES data in comparison to the baseline model. Performance gain by the introduction of Energy SES data is negligible and *MAPE* for *Lasso* and *ElasticNet* is much worse.

Table 7 **Model performance and hyperparameters for Energy SES data.**

Model	RMSE	R^2	MAPE	Alpha	L_1 Proportion
Baseline	2.1686	0.8844	17.082	-	-
Linear Regression	2.1768	0.8835	17.51	-	-
Lasso Regression	2.1656	0.8847	101.455	0.007	-
Ridge Regression	2.1728	0.884	17.453	0.316	-
Elastic Net	2.1651	0.8848	101.092	0.006	0.99

Table 8 shows the feature coefficients for the models based on Energy SES data. What is interesting to see is that for the *Lasso* and *ElasticNet* models, the regularization term removes all the features except for the *Distance* feature.

When considering the paper by Ohsugi and Koshizuka (2018), an expectation can be that higher energy and gas usage leads to lower delivery times. However, the data does not show a similar trend for all the building and owner types. It could be that energy and gas usage is influenced by further characteristics, such as building age, interfering with the occupancy effect of having higher energy and gas usage when there is someone present, which would in turn mean lower delivery times.

Table 8 **Feature coefficients for models trained on Energy SES data.**

Feature	Linear	Lasso	Ridge	ElasticNet
Distance	46.8751	46.4191	46.5102	46.3112
Avg_energy_usage	-0.0660	-	0.0890	-
Avg_energy_usage_apps	-0.2762	-	-0.1875	-
Avg_energy_usage_terraced	1.3586	-	1.1810	-
Avg_energy_usage_corner	-0.6275	-	-0.5822	-
Avg_energy_usage_rental	1.5060	-	1.0971	-

Feature	Linear	Lasso	Ridge	ElasticNet
Avg_energy_usage_owner_occupied	-2.0326	-	-1.6161	-
Avg_gas_usage	2.4553	-	1.6426	-
Avg_gas_usage_terraced	1.5240	-	1.4268	-
Avg_gas_usage_corner	-0.7099	-	-0.5156	-
Avg_gas_usage_rental	-2.8196	-	-2.1413	-
Avg_gas_usage_owner_occupied	-0.4205	-	-0.4007	-

Table 9 shows the performance of the Energy SES data on the baseline residuals. Again, there is no proof for Energy SES data improving prediction further than with the use of the *Distance* feature as it is unable to explain the variance in the residuals.

Table 9 **R^2 values for models predicting residuals of the baseline model on Energy SES data.**

	Linear	Lasso	Ridge	ElasticNet
R^2	-0.007	0.000	-0.005	0.000

5.2.4 Education model

Table 10 shows the performance and hyperparameters of the models based on Education SES data in comparison to the baseline model. *Lasso* and *ElasticNet* models perform slightly better in terms of *RMSE* and R^2 compared to the baseline, though a lot worse when considering *MAPE*. The high L_1 proportion for *ElasticNet* also indicates that the model improves from removing features.

Table 10 **Model performance and hyperparameters for Education SES data.**

Model	RMSE	R^2	MAPE	Alpha	L_1 Proportion
Baseline	2.1686	0.8844	17.082	-	-
Linear Regression	2.1768	0.8835	17.51	-	-
Lasso Regression	2.1656	0.8847	101.455	0.006	-

Model	RMSE	R^2	MAPE	Alpha	L_1 Proportion
Ridge Regression	2.1728	0.884	17.453	0.058	-
Elastic Net	2.1651	0.8848	101.092	0.004	0.99

The feature coefficients in *Table 11* show a similar picture where the *Distance* feature is most influential, the other feature coefficients have a much smaller impact. The *Lasso* and *ElasticNet* models remove the low and medium education level from the model. An increase in low and medium education level proportions has a negative effect on prediction, while high education level proportions have a positive effect on prediction. This effect can be expected as highly educated individuals have a higher employment rate (Centraal Bureau voor de Statistiek, 2020), which could mean that they are home less often.

Table 11 **Feature coefficients for models trained on Education SES data.**

Feature	Linear	Lasso	Ridge	ElasticNet
Distance	46.8678	46.4823	46.8008	46.4948
Percent_edulevel_low	-0.0645	-	-0.0638	-
Percent_edulevel_med	-0.2227	-	-0.2194	-
Percent_edulevel_high	0.0946	0.0275	0.0959	0.0986

When using the Education SES data to predict the residuals of the baseline model, again there are no indications that it can significantly explain the variance of the residual as can be seen in *Table 12*.

Table 12 **R^2 values for models predicting residuals of the baseline model on Education SES data.**

	Linear	Lasso	Ridge	ElasticNet
R^2	-0.001	0.000	-0.001	0.000

5.2.5 Labor model

Table 13 shows the performance and hyperparameters of the models based on Labor SES data in comparison to the baseline model. Like the previous models there is no indication that the models that include SES data perform substantially better than the baseline model.

Table 13 **Model performance and hyperparameters for Labor SES data.**

Model	RMSE	R^2	MAPE	Alpha	L_1 Proportion
Baseline	2.1686	0.8844	17.082	-	-
Linear Regression	2.1686	0.8844	17.075	-	-
Lasso Regression	2.1663	0.8846	101.883	0.005	-
Ridge Regression	2.168	0.8845	17.0853	0.058	-
Elastic Net	2.166	0.8847	101.707	0.004	0.99

Table 14 shows the feature coefficients. *Lasso* and *ElasticNet*'s regularization terms remove the features other than *Distance* from the model. *Linear* and *Ridge* models show a very small negative and positive effect, for the *employed* and *employee* proportion, respectively.

Table 14 **Feature coefficients for models trained on Labor SES data.**

Feature	Linear	Lasso	Ridge	ElasticNet
Distance	46.8582	46.5446	46.7913	46.4932
Percent_employed	-0.0922	-	-0.0921	-
Percent_employees	0.0958	-	0.0957	-

The features in the Labor SES data are unable to explain any of the variance in the residuals from the baseline model.

Table 15 **R^2 values for models predicting residuals of the baseline model on Labor SES data.**

	Linear	Lasso	Ridge	ElasticNet
R^2	0.000	0.000	0.000	0.000

5.2.6 Social Security model

Table 16 shows the performance and hyperparameters of the models based on Social Security SES data in comparison to the baseline model. There is no indication of a significant improvement in prediction when including the Social Security SES data into the models.

Table 16 **Model performance and hyperparameters for Social Security SES data.**

Model	RMSE	R^2	MAPE	Alpha	L_1 Proportion
Baseline	2.1686	0.8844	17.082	-	-
Linear Regression	2.1705	0.8842	17.117	-	-
Lasso Regression	2.1659	0.8847	101.668	0.006	-
Ridge Regression	2.1696	0.8843	17.131	0.087	-
Elastic Net	2.1651	0.8848	101.092	0.006	0.99

Table 17 confirms that *Lasso* and *ElasticNet* models do not improve with the inclusion of Social Security SES data. *Bijstand* and *WW* show negative effects and *AO* and *AOW* positive effects on delivery time prediction for the *Linear* and *Ridge* model. The *Percent_AO* feature shows the highest impact next to *Distance*.

Table 17 **Feature coefficients for models trained on Social Security SES data.**

Feature	Linear	Lasso	Ridge	ElasticNet
Distance	46.8051	46.4818	46.7055	46.3112
Percent_bijstand	-0.3536	-	-0.3495	-

Feature	Linear	Lasso	Ridge	ElasticNet
Percent_AO	2.5920	-	2.5559	-
Percent_WW	-0.1156	-	-0.1168	-
Percent_AOW	0.0220	-	0.0241	-

Table 18 shows that the Social Security SES data is unable to explain the variance in the residuals of the baseline model.

Table 18 **R^2 values for models predicting residuals of the baseline model on Social Security SES data.**

	Linear	Lasso	Ridge	ElasticNet
R^2	-0.0022	-0.0001	-0.0017	-0.0001

5.2.7 Care model

Table 19 shows the performance and hyperparameters of the models based on Care SES data in comparison to the baseline model. *WMO_clients(per 1000)* is not used because it contains the same data as *Percent_WMO_clients*. The model performance metrics and the hyperparameters do not show significant improvement for models that use Care SES data in their prediction.

Table 19 **Model performance and hyperparameters for Care SES data.**

Model	RMSE	R^2	MAPE	Alpha	L_1 Proportion
Baseline	2.1686	0.8844	17.082	-	-
Linear Regression	2.1712	0.8841	17.12	-	-
Lasso Regression	2.1663	0.8846	101.883	0.005	-
Ridge Regression	2.1698	0.8843	17.159	0.172	-
Elastic Net	2.1655	0.8847	101.4	0.005	0.99

Table 20 shows that the *Lasso* and *ElasticNet* models completely remove the Care SES data from the model. The different types of *Youth Services* show a negative and positive effect for *Linear* and *Ridge* models with *Percent_youth_services* showing the largest impact. Then percentage of *WMO* clients has a very small positive effect.

***Table 20* Feature coefficients for models trained on Care SES data.**

Feature	Linear	Lasso	Ridge	ElasticNet
Distance	46.8186	46.5446	46.6231	46.4021
Percent_youth_services(natura)	-0.8540	-	-0.3281	-
Percent_youth_services	2.1452	-	1.7606	-
Percent_WMO_clients	0.0525	-	0.1287	-

Table 21 shows that Care SES data is unable to explain any of the variance in the residuals of the baseline model.

***Table 21* R^2 values for models predicting residuals of the baseline model on Care SES data.**

	Linear	Lasso	Ridge	ElasticNet
R^2	-0.0028	-0.0001	-0.0026	-0.0001

5.2.8 Business Locations model

Table 22 shows the performance and hyperparameters of the models based on Business Locations SES data in comparison to the baseline model. While *Lasso* and *ElasticNet* regression perform slightly better in terms of *RMSE* and R^2, their *MAPE* performance is much worse. There is no indication that the inclusion of Business Locations SES data in the regression models improves prediction.

Table 22 **Model performance and hyperparameters for Business Locations SES data.**

Model	RMSE	R^2	MAPE	Alpha	L_1 Proportion
Baseline	2.1686	0.8844	17.082	-	-
Linear Regression	2.1791	0.8833	17.471	-	-
Lasso Regression	2.1661	0.8847	101.528	0.007	-
Ridge Regression	2.1749	0.8837	17.391	0.183	-
Elastic Net	2.1680	0.8845	102.264	0.003	0.99

The feature coefficients in *Table 23* show that the *Lasso* model has removed all but one coefficient (*Businesses_per_inhabitant)* from the model which has a negative effect. The *ElasticNet* has retained one more feature on *Cultural* and *Recreation* businesses that also has a negative effect. For the *Linear* and *Ridge* models, the *Distance, agricultural*, *trade* and *service* feature have a positive effect, the rest has a negative effect when the proportions increase. *Service_businesses_per_inhabitant* shows the largest impact followed by *Transport_IT_businesses_per_inhabitant* and *Cultural_recreation_businesses_per_inhabitant*.

Table 23 **Feature coefficients for models trained on Business Locations SES data.**

Feature	Linear	Lasso	Ridge	ElasticNet
Distance	46.8506	46.4404	46.6535	46.6115
Businesses_per_inhabitant	-0.5646	-0.2597	-0.5659	-0.4464
Agricultural_businesses_per_inhabitant	0.6188	-	0.4058	-
Industry_businesses_per_inhabitant	-0.8468	-	-0.5821	-
Trade_catering_businesses_per_inhabitant	0.4379	-	0.2421	-
Transport_IT_businesses_per_inhabitant	-4.8273	-	-2.2099	-
Finance_businesses_per_inhabitant	-0.8088	-	-0.3547	-
Service_businesses_per_inhabitant	7.9490	-	4.7874	-
Cultural_recreation_businesses_per_inhabitant	-2.5239	-	-2.3064	-0.5236

The R^2 metrics for the prediction of the residuals of the baseline model with models based on the Business Locations SES data indicate that they are unable to explain the variance in the residuals.

Table 24 **R^2 values for models predicting residuals of the baseline model on Business Locations SES data.**

	Linear	**Lasso**	**Ridge**	**ElasticNet**
R^2	-0.0098	-0.0001	-0.0049	-0.0001

5.2.9 Motor Vehicles model

Table 25 shows the performance and hyperparameters of the models based on Motor Vehicles SES data in comparison to the baseline model. The metrics show no sign of significant improvement in terms of *RMSE,* R^2 and *MAPE* for any of the models.

Table 25 **Model performance and hyperparameters for Living SES data.**

Model	**RMSE**	**R^2**	**MAPE**	**Alpha**	**L_1 Proportion**
Baseline	2.1686	0.8844	17.082	-	-
Linear Regression	2.1704	0.8842	17.457	-	-
Lasso Regression	2.167	0.8846	103.167	0.003	-
Ridge Regression	2.1699	0.8843	17.492	0.083	-
Elastic Net	2.1664	0.8846	102.855	0.003	0.99

The feature coefficients for the models trained with Motor Vehicles SES data are displayed in *Table 26*. The *Lasso* and *ElasticNet* models have reatined the features *Avg_other_cars* and *Avg-_motorbikes*, which are the most impactful as well, in addition to the *Distance* feature. The number of cars per inhabitant, household and per square kilometer have a positive effect on delivery time prediction as does the average number of motorbikes. The other two features have a negative effect on prediction.

Table 26 **Feature coefficients for models trained on Motor Vehicles SES data.**

Feature	Linear	Lasso	Ridge	ElasticNet
Distance	46.8365	46.6487	46.7419	46.5625
Avg_cars	7.6868	-	0.2979	-
Avg_petrol_cars	-5.5431	-	-0.1146	-
Avg_other_cars	-4.1623	-1.1606	-1.6678	-1.1557
Cars_per_household	0.0122	-	0.0115	-
Cars_per_sqkm	0.0755	-	0.0770	-
Avg_motorbikes	1.5591	1.2773	1.5261	1.2805

Table 27 shows that models trained on the Motor vehicles SES data are unable to explain the variance in the residuals of the baseline model.

Table 27 **R^2 values for models predicting residuals of the baseline model on Motor vehicles SES data.**

	Linear	Lasso	Ridge	ElasticNet
R^2	-0.0019	-0.0002	-0.0018	-0.0002

5.2.10 Services model

Table 28 shows the performance and hyperparameters of the models based on Services SES data in comparison to the baseline model. The baseline model outperforms all the other models in terms of *RMSE*, R^2 and *MAPE*.

Table 28 **Model performance and hyperparameters for Services SES data.**

Model	RMSE	R^2	MAPE	Alpha	L_1 Proportion
Baseline	2.1686	0.8844	17.082	-	-

Model	RMSE	R^2	MAPE	Alpha	L_1 Proportion
Linear Regression	2.1729	0.8839	17.177	-	-
Lasso Regression	2.169	0.8844	101.954	0.006	-
Ridge Regression	2.1725	0.884	17.184	0.051	-
Elastic Net	2.1689	0.8844	101.727	0.005	0.99

The *Lasso* and *ElasticNet* models retain the *Avg_distance_to_gp(km)* and *Nr_of_schools_within_3km* features. For all the models, the *Distance* feature is the most influential and the effects of the other features are negligible.

Table 29 **Feature coefficients for models trained on Services SES data.**

Feature	Linear	Lasso	Ridge	ElasticNet
Distance	46.8043	46.4433	46.7449	46.3595
Avg_distance_to_gp(km)	0.59817	0.2836	0.59646	0.33969
Avg_distance_to_large_supermarket(km)	0.05809	-	0.05796	-
Avg_distance_to_daycare(km)	0.03229	-	0.0304	-
Avg_distance_to_school(km)	-0.0815	-	-0.075	-
Nr_of_schools_within_3km	-0.2457	-0.2183	-0.2476	-0.2292

The R^2 scores in *Table 30* suggests that the features in the Services SES data are unable to explain any of the variance in the residuals of the baseline model.

Table 30 **R^2 values for models predicting residuals of the baseline model on Services SES data.**

	Linear	Lasso	Ridge	ElasticNet
R^2	-0.0044	-0.0001	-0.0044	-0.0006

5.2.11 Surface model

Table 31 shows the performance and hyperparameters of the models based on Surface SES data in comparison to the baseline model. The baseline model outperforms all the other models in terms of *RMSE, R^2* and *MAPE*.

Table 31 **Model performance and hyperparameters for Surface SES data.**

Model	RMSE	R^2	MAPE	Alpha	L_1 Proportion
Baseline	2.1686	0.8844	17.082	-	-
Linear Regression	2.1858	0.8826	17.165	-	-
Lasso Regression	2.1804	0.8831	103.241	0.002	-
Ridge Regression	2.1850	0.8826	17.130	0.075	-
Elastic Net	2.1801	0.8832	103.04	0.002	0.99

The feature coefficients in *Table 32* show that *Lasso* removes all but the *Surface_land(ha)* feature and *ElasticNet* includes *Surface(ha)* as well. In the *Linear* model, the coefficients for *Surface(ha)* and *Surface_land(ha)* are large, though as they are strongly correlated (Total area and Total area excluding surface water), they cancel each other out.

Table 32 **Feature coefficients for models trained on Surface SES data.**

Feature	Linear	Lasso	Ridge	ElasticNet
Distance	46.6009	46.4965	46.4857	46.4381
Surface(ha)	-182.2419	-	1.6168	0.3435
Surface_land(ha)	181.0141	2.8333	1.9761	2.5106
Surface_water(ha)	4.4355	-	-0.2531	-

The R^2 scores in *Table 33* suggests that the features in the Surface SES data are unable to explain any of the variance in the residuals of the baseline model.

Table 33 **R^2 values for models predicting residuals of the baseline model on Surface SES data.**

	Linear	Lasso	Ridge	ElasticNet
R^2	-0.0175	-0.0001	-0.0158	-0.0025

5.2.12 Urbanity model

Table 23 shows the performance and hyperparameters of the models based on Urbanity SES data in comparison to the baseline model. The baseline model outperforms all the other models in terms of *RMSE, R^2* and *MAPE*.

Table 34 **Model performance and hyperparameters for Urbanity SES data.**

Model	RMSE	R^2	MAPE	Alpha	L_1 Proportion
Baseline	2.1686	0.8844	17.082	-	-
Linear Regression	2.1762	0.8836	17.111	-	-
Lasso Regression	2.1738	0.8838	103.138	0.003	-
Ridge Regression	2.176	0.8836	17.116	0.033	-
Elastic Net	2.1734	0.8839	102.843	0.003	0.99

The feature coefficients in *Table 35* show that urbanity has a negative effect on parcel delivery time estimation. When a neighborhood has a higher urbanity, the estimated delivery time estimate goes down. The *Lasso* and *ElasticNet* model both remove the *Degree_of_urbanity* feature. This could be because this is a categorical variable while the *Urbanity(sqkm)* is a continuous variable which could allow for better prediction. The impact of both features is however minimal.

Table 35 **Feature coefficients for models trained on Urbanity SES data.**

Feature	**Linear**	**Lasso**	**Ridge**	**ElasticNet**
Distance	46.7813	46.5973	46.7429	46.5108
Degree_of_urbanity	-0.1520	-	-0.1493	-
Urbanity(sqkm)	-0.6405	-0.4545	-0.6402	-0.4597

The R^2 scores in *Table 36* suggests that the features in the Surface SES data are unable to explain any of the variance in the residuals of the baseline model.

Table 36 **R^2 values for models predicting residuals of the baseline model on Urbanity SES data.**

	Linear	**Lasso**	**Ridge**	**ElasticNet**
R^2	-0.0077	-0.0001	-0.0076	-0.0027

5.2 Models with *Log_Distance*

To rule out that the models do not perform well because the untransformed *Distance* value is used that was leptokurtic and skewed, it is important to consider the performance of models trained on the *Log_Distance* feature and the SES data in comparison to a baseline model that only uses *Log_Distance* as a predictor. *Table 37* shows the performance metrics for the best performing algorithm per SES data category. From this table it becomes clear that using *Log_Distance* for prediction does not affect model performance.

Table 37 **Performance metrics of best performing models per SES data category trained on the *Log_Distance* feature.**

Model	Algorithm	RMSE	R^2	MAPE
Baseline	Linear	3.590	0.7028	52.880
Population	Lasso	3.4934	0.7000	225.751
Living	Lasso	3.5283	0.6940	246.235
Energy	Lasso	3.5288	0.6939	328.711
Education	Linear	3.5278	0.6940	85.108
Labor	Lasso	3.5286	0.6939	503.327
Social Security	Lasso	3.5288	0.6939	328.711
Care	Lasso	3.5288	0.6939	328.711
Business Locations	Lasso	3.5259	0.6944	492.525
Motor Vehicles	Linear	3.5090	0.6973	65.475
Services	Linear	3.5127	0.6967	68.734
Surface	Lasso	3.5449	0.6911	336.386
Urbanity	Linear	3.5163	0.6960	100.702

The performance of the SES data models on the residuals of the *Log_Distance* baseline in *Table 38* also show that they are unfit for predicting the variance in the residuals of the baseline *Log_Distance* model. While some of the R^2 scores are above zero, it is nowhere near an acceptable value for a well explaining R^2 value.

Table 38 R^2 scores for *Log_Distance* models

Model	Algorithm	R^2
Population	Lasso	0.0197
Living	Lasso	0.0003
Energy	Lasso & ElasticNet	-0.00002
Education	Linear & Ridge	0.0001
Labor	Lasso & ElasticNet	-0.00002
Social Security	Lasso & ElasticNet	-0.00002
Care	Lasso & ElasticNet	-0.00002
Business Locations	ElasticNet	0.002
Motor Vehicles	Linear	0.0107
Services	Linear	0.0086
Surface	Lasso	-0.0034
Urbanity	ElasticNet	0.0071

6. Discussion

The performance of the models that included SES data did not significantly improve compared to the baseline model, regardless of transforming the *Distance* feature. Also, for the models trained on the residuals of the baseline model, there is no indication that SES data influences prediction.

While the models that are trained on the SES data do show some trends that lend for causal analysis, for example in neighborhoods where with an increase in the proportion of inhabitants that is older than 65 years the predicted delivery time goes down, these trends could be random and might only exist in the TDV dataset reducing ecological validity and making causal analysis hard to perform and justify.

The feature that is most important in parcel delivery time estimation is *Distance*. The similar performance of the baseline model to the models that include SES data confirms this. When it comes to SES data, the only SES data category that showed improvement for all the tested regression and regularization techniques in terms of *RMSE* and R^2 was the Population SES data in combination with the *Distance* feature. The model that performed best on the Population SES data was the *Ridge* model (Baseline model: *RMSE* = 2.1686, R^2 = 0.8844; *Ridge: RMSE* = 2.1646, R^2 = 0.8851) which constituted to a decrease of 0.31% in *RMSE* and a 0.0007 increase in *R2.* It is therefore safe to say that the impact of SES data on prediction is negligible, also as the *MAPE* was lower for all the models regardless of the type of SES data. These marginal improvements in *RMSE* and R^2 do not justify concluding that the Population SES data or any of the other SES data categories used in the models have a significant impact on prediction.

Considering algorithm performance, it is hard to say what algorithm proved best. Essentially none of the algorithms proved much better than the baseline, selecting the best algorithm would therefore not justify the observed results.

6.1 Limitations

The lack of improvement in the SES data models does not mean that there is no relation between the categories of SES data used for modeling and parcel delivery time prediction. It could be that the trends that show in the feature coefficients hold some truth, or there are other trends that were not uncovered by the methodologies applied.

6.1.1 Algorithm selection

Because one of the goals for this project was to obtain a level of ecological validity, the algorithms that were selected for the modelling are not the most advanced ones. The trade-off between prediction and transparency might have been oriented too much towards transparency. This led to models that, in theory predict well, but are unable to pick up on trends that are present in the SES data as they rely too heavily on the *Distance* feature. Because the algorithms used are based on linear regression, it might be interesting to see if there are other models that can pick up trends in the data that are not linear, though this might come at the cost of transparency.

6.1.2. The TDV dataset

The first, roughly cleaned, dataset consisted of 20005 instances. After cleaning was done 3635 instances remained which is a decrease of 81.83%. This indicates that a lot of the data was not deemed as appropriate and raises the question to what extent the remaining data is valid. The large reduction in size could be due to the lack of verification of the order completion location and the delivery location. If a dataset were to be used that contained such data, it would be much more efficient in filtering out invalid instances. The filtering technique used, selecting a boundary for average speed, is far from ideal.

6.1.3 COVID-19

Another factor that might have played a role is the global COVID-19 pandemic. The TDV dataset contains data from the time where (lockdown) measures taken by the Dutch government to fight the pandemic meant that a lot of people were working from home and shops and catering businesses were closed. This could result in trends specific to this situation that are not generalizable to the situation after the pandemic. Also, the CBS dataset used originates from 2019 as the 2020 version was not yet complete, this could also lead to a disparity between the datasets. An option could have been to swap the features that were complete in the 2020 dataset into the 2019 dataset, though this this could potentially harm the effects that showed in the data.

7. Conclusion

While the expected increase in population in urban areas and the increase in urban freight operations that is associated with this increase require innovative solutions for urban freight transport, this thesis was unable to deliver a meaningful contribution to this field other than that, for this dataset, there is no apparent indication that SES impacts parcel delivery. It could well be that there is no relation between SES and delivery, but it could also be that the data quality was not high enough or the applied methodologies were unable to pick up on trends that were present.

This also means that the models that were built for this thesis project have no ecological validity to, for example, select locations for new UMCs or to measure the impact that urban freight transport has on different neighborhoods in the city of Eindhoven. The impact of SES data on delivery is thus, regardless of category, negligible.

7.1 Future work

While this thesis was unable to show an effect between SES data and delivery time prediction, there are some aspects that can be improved in the research design. By eliminating some of the uncertainties surrounding data quality or the selection of a different set of algorithms, for example, another project might be able to show that there are effects between SES data and delivery time prediction or support the findings of this thesis that there are no significant effects.

The SES data that was used for this thesis is quite broad as well, as briefly mentioned before, the energy data might be influenced by average building age which could impact prediction effectiveness. As this thesis deployed a broad and exploratory perspective, it might be worthwhile to investigate improving the SES data as well. For example, by including energy efficiency labels for houses typical to a neighborhood in prediction modelling.

Finally, the dataset that was used for this thesis is relatively small and the effects seemed to be very subtle if present at all. What would be an interesting opportunity is to see what happens when a similar dataset from a large company such as PostNL, DHL or DPD is used with the same goal. Because these datasets are much larger, it might be possible to discover the small effects that SES data can have.

References

Archetti, C., & Bertazzi, L. (2020). Recent challenges in Routing and Inventory Routing: E-commerce and last-mile delivery. *Networks, 77*(2), 255-268. doi:https://doi.org/10.1002/net.21995

Bektas, T., Crainic, T. G., & van Woensel, T. (2015). From Managing Urban Freight to Smart City Logistics Networks. *CIRRELT, 17*, 2.

Centraal Bureau voor de Statistiek. (2020). *Jaarrapport Integratie 2020.* Den Haag: Centraal Bureau voor de Statistiek.

Chu, H., Zhang, W., Bai, P., & Chen, Y. (2021). Data-driven optimization for last-mile delivery. *Complex & Intelligent Systems*, 1-14. doi:https://doi.org/10.1007/s40747-021-00293-1

Conway, A., Fatisson, P.-E., Eickemeyer, P., Cheng, J., & Peters, D. (2012). URBAN MICRO-CONSOLIDATION AND LAST MILE GOODS DELIVERY BY FREIGHT-TRICYCLE IN MANHATTAN: OPPORTUNITIES AND CHALLENGES. *Transportation Research Board, 91*, 1-17.

Cruz de Araujo, A., & Etemad, A. (2021). End-to-End Prediction of Parcel Delivery Time with Deepl Learning for Smart-City Applications. *IEEE Internet of Things Journal*, 1-14. doi:10.1109/JIOT.2021.3077007

Daoud, A., Kim, R., & Subramanian, S. V. (2019). Predicting women's height from their socioeconomic status: A machine learning approach. *Social Science & Medicine, 238*, 1-9. doi:https://doi.org/10.1016/j.socscimed.2019.112486

Davydenko, A., & Fildes, R. (2016). *Forecast Error Measures: Critical Review and Practical Recommendations.* John Wiley & Sons Inc.

Gemeente Eindhoven. (2021, 5 15). *OP WEG NAAR EEN NUL-EMISSIEZONE*. Retrieved from Eindhoven.nl: https://www.eindhoven.nl/projecten/nul-emissiezone/op-weg-naar-een-nul-emissiezone

Gevaers, R., Van de Voorde, E., & Vanelslander, T. (2009). CHARACTERISTICS OF INNOVATIONS IN LAST MILE LOGISTICS - USING BEST PRACTICES, CASE STUDIES AND MAING THE LINK WITH GREEN AND SUSTAINABLE LOGISTICS-. *Association for European Transport and contributors*, 1-21.

Golden, B., Raghavan, S., & Wasil, E. (Eds.). (2008). *The Vehicle Routing Problem: Latest Advances and New Challenges.* New York ; London: Springer. doi:10.1007/978-0-387-77778-8

Hagen, T., & Scheel-Kopeinig, S. (2021). Would customers be willing to use an alternative (chargeable) delivery concept for the last mile? *Research in Transportation Business & Management*, 1-13. doi:https://doi.org/10.1016/j.rtbm.2021.100626.

Hughes, S., Moreno, S., Yushimito, W. F., & Huerta-Cánepa, G. (2019). Evaluation of machine learning methodologies to predict stop delivery times from GPS data. *Transportation Research Part C: Emerging Technologies, 109*, 289-304. doi:https://doi.org/10.1016/j.trc.2019.10.018

McLaren, L. (2007). Socioeconomic Status and Obesity. *Epidemiologic REviews, 29*(1), 29-48. doi:https://doi.org/10.1093/epirev/mxm001

Muñuzuri, J., Cortés, P., Grosso, R., & Guadix, J. (2012). Selecting the location of minihubs for freight delivery in congested downtown areas. *Journal of Computational Science, 3*, 228-237. doi:https://doi.org/10.1016/j.jocs.2011.12.002

Naumov, V., Vasiutina, H., & Solarz, A. (2021). Modelling demand for deliveries by cargo bicycles in the Old Town of Kraków. *Transportation Research Procedia, 52*, 11-18. doi:https://doi.org/10.1016/j.trpro.2021.01.003

Ohsugi, S., & Koshizuka, N. (2018). Delivery route optimization through occupancy prediction from electricity usage. *IEEE International Conference on Computer Software & Application, 42*, 1-8.

Schultz, W. M., Kelli, H. M., Lisko, J. C., Vargheses, T., Shen, J., Sandesara, P., . . . Sperling, L. S. (2018). Socioeconomic Status and Cardiovascular Outcomes. *Circulation, 137*(20), 2166-2178. doi:https://doi.org/10.1161/CIRCULATIONAHA.117.029652

Slabinac, M. (2015). INNOVATIVE SOLUTIONS FOR A "LAST-MILE" DELIVERY - A EUROPEAN EXPERIENCE. *15th international scientific conference Business Logistics in Modern Management*, 111-129.

Topsector Logistiek. (2019). *Laadinfrastructuur voor elektrische voertuigen in stadlogistiek.* Connekt.

World Health Organization. (2005). *Health effects of transport-related air pollution.* Copenhagen: WHO Regional Office for Europe.

Wrighton, S., & Reiter, K. (2016). CycleLogistics - moving Europe forward! *Transportation Research Procedia, 12*, 950-958.

Appendix A

Date	Customer	pick-up street	pick-up nr	pick-up postcode	Pick-up place	Delivery street	Delivery nr	Delivery postcode	Delivery place	Status	Pick-up at	Delivery at	Messenger
5-11-2020	1	Frederiklaan	10	5616 NH	Eindhoven	Stadhuisplein	1	5611 EM	Eindhoven	Delivered	19:13:00	20:22:00	Sjors
5-11-2020	12	-	-	-	-	Stratumseind	63	5611 ET	Eindhoven	Delivered	-	20:25:00	Sjors
23-2-2021	76	Luchthavenweg	25	5657 EA	Eindhoven	Hallenweg	1	5615 PP	Eindhoven	Delivered	10:15:00	11:30:00	Drew
23-2-2021	54	Hallenweg	1	5615 PP	Eindhoven	Luchthavenweg	67	5657 EA	Eindhoven	Delivered	08:15:00	11:34:00	Drew

Appendix B – Features in the dataset

Feature name	# Missing	Description
Postcode	0	Postal code
Huisnummer	0	House number
Straat	0	Street name
Plaats	0	City name
Distance	0	Distance travelled to destination
Koerier	0	Messenger
Inhabitants	0	Number of inhabitants
Percent_men	106	The proportion of male inhabitants
Percent_women	106	The proportion of female inhabitants
Percent_0-15age	106	The proportion of inhabitants aged between 0 and 15 years
Percent_15-25age	106	The proportion of inhabitants aged between 15 and 25 years
Percent_25-45age	106	The proportion of inhabitants aged between 25 and 45 years
Percent_45-65age	106	The proportion of inhabitants aged between 45 and 65 years
Percent_65+age	106	The proportion of inhabitants older than 65 years
Percent_not_married	106	The proportion of inhabitants that are not married
Percent_widowed	106	The proportion of inhabitants that are widowed
Percent_migration_western	106	The proportion of inhabitants with a western migration background
Percent_migration_non_western	106	The proportion of inhabitants with a non-western migration background
Percent_migration_non_western(Morocco)	106	The proportion of inhabitants with a Moroccan migration background

Feature name	# Missing	Description
Percent_migration_non_western(Antilles)	106	The proportion of inhabitants with an Antillean migration background
Percent_migration_non_western(Suriname)	106	The proportion of inhabitants with a Surinam migration background
Percent_migration_non_western(Turkije)	106	The proportion of inhabitants with a Turkish migration background
Percent_migration_non_western(Others)	106	The proportion of inhabitants with a non-western migration background other than the aforementioned
Births(per-1000)	0	The number of births per 1000 inhabitants
Deaths(per-1000)	0	The number of deaths per 1000 inhabitants
Households	0	The number of households
Percent_1person_hh	110	The proportion of 1 person households
Percent_no_kids_hh	110	The proportion of households that do not have children
Percent_w_kids_hh	110	The proportion of households that have children
Avg_size_hh	110	The average size of a household
Population_density_sqkm	0	The number of inhabitants per square kilometer
Housing_stock_per_inhabitant	106	The number of homes per inhabitant
Avg_price_home(x1000)	424	The average price of a home in euros (x1000)
Percent_1family_housing	385	The proportion of 1 family homes
Percent_multiple_family_housing	385	The proportion of multiple family homes
Percent_inhabited	385	The proportion of homes that are inhabited
Percent_uninhabited	385	The proportion of homes that are uninhabited
Percent_owner_inhabited	385	The proportion of homes that is owner inhabited
Percent_rental_properties	385	The proportion of rental homes

Feature name	# Missing	Description
Percent_housing_corporation_rental_properties	385	The proportion of rental homes that are owned by housing corporations
Percent_rental_properties_other_owners	385	The proportion of rental homes that are owned owners other than housing corporations
Percent_owner_unknown	385	The proportion of homes for which the owner is unknown
Percent_homes_build_before_2000	385	The proportion of homes built before 2000
Percent_homes_build_after_2000	385	The proportion of homes built after 2000
Avg_energy_usage (kWh)	177	The average energy usage of a home
Avg_energy_usage_apps (kWh)	671	The average energy usage of an appartement
Avg_energy_usage_terraced (kWh)	540	The average energy usage of a terraced home
Avg_energy_usage_corner (kWh)	622	The average energy usage of a corner home
Avg_energy_usage_semidetached (kWh)	964	The average energy usage of a semi-detached home
Avg_energy_usage_detached (kWh)	1614	The average energy usage of a detached home
Avg_energy_usage_rental (kWh)	442	The average energy usage of a rental home
Avg_energy_usage_owner_occupied (kWh)	256	The average energy usage of an owner occupied home
Avg_gas_usage (m^3)	410	The average gas usage of a home
Avg_gas_usage_apps (m^3)	974	The average gas usage of an appartement
Avg_gas_usage_terraced (m^3)	721	The average gas usage of a terraced home
Avg_gas_usage_corner (m^3)	745	The average gas usage of a corner home
Avg_gas_usage_semidetached (m^3)	980	The average gas usage of a semi-detached home
Avg_gas_usage_detached (m^3)	1610	The average gas usage of a detached home
Avg_gas_usage_rental (m^3)	574	The average gas usage of a rental home
Avg_gas_usage_owner_occupied (m^3)	347	The average gas usage of a owner occupied home

Feature name	# Missing	Description
Percent_district_heating	3728	The proportion of homes that are connected to district heating
Percent_edulevel_low	294	The proportion of the inhabitants with a low educational level
Percent_edulevel_med	294	The proportion of inhabitants with a medium educational level
Percent_edulevel_high	198	The proportion of inhabitants with a high educational level
Percent_employed	428	The proportion of inhabitants that are employed
Percent_employees	454	The proportion of working inhabitants that are employees
Percent_employers	454	The proportion of working inhabitants that are employers
Percent_bijstand	168	The proportion of inhabitants that receive benefits
Percent_AO	168	The proportion of inhabitants that are incapacitated for work
Percent_WW	168	The proportion of inhabitants that receive unemployment benefits
Percent_AOW	168	The proportion of inhabitants that receive social security
Percent_youth_services(natura)	526	The proportion of inhabitants that receive youth services in natura
Percent_youth_services	526	The proportion of inhabitants that receive youth services
Percent_WMO_clients	546	The proportion of inhabitants that receive benefits from the social support act
WMO_clients(per 1000)	546	The number of inhabitants that receive benefits from the social support act per 1000
Businesses_per_inhabitant	106	The number of businesses per inhabitant
Agricultural_businesses_per_inhabitant	145	The number of agricultural businesses per inhabitant
Industry_businesses_per_inhabitant	145	The number of industrial businesses per inhabitant
Trade_catering_businesses_per_inhabitant	145	The number of trade and catering businesses per inhabitant
Transport_IT_businesses_per_inhabitant	145	The number of IT and transport business per inhabitant

Feature name	# Missing	Description
Finance_businesses_per_inhabitant	145	The number of finance businesses per inhabitant
Service_businesses_per_inhabitant	145	The number of service businesses per inhabitant
Cultural_recreation_businesses_per_inhabitant	145	The number of cultural and recreational businesses per inhabitant
Avg_cars	106	The number of cars per inhabitant
Avg_petrol_cars	106	The number of petrol cars per inhabitant
Avg_other_cars	106	The number of non-petrol cars per inhabitant
Cars_per_household	415	The number of cars per household
Cars_per_sqkm	415	The number of cars per square kilometer
Avg_motorbikes	106	The number of motorbikes per inhabitant
Avg_distance_to_gp(km)	115	The average distance to a general practitioner
Avg_distance_to_large_supermarket(km)	115	The average distance to a large supermarket
Avg_distance_to_daycare(km)	115	The average distance to a daycare facility
Avg_distance_to_school(km)	115	The average distance to a school
Nr_of_schools_within_3km	115	The number of schools within a 3 km radius
Surface(ha)	0	The surface of the neighborhood
Surface_land(ha)	0	The area of the surface that is land
Surface_water(ha)	0	The area of the surface that is water
Most_common_pc	0	The most common postal code for the neighborhood
Pc_coverage	0	The postal code coverage 1: > 90% same postal code, 2: 81-90% 3: 71-80% “ “ “ 4: 61-70% “ “ “ 5: 51-60% “ “ “ 6 < 50% “ “ “

Feature name	# Missing	Description
Degree_of_urbanity	0	Degree of urbanity: 1: >= 2500 addresses per square kilometer 2: 1500 – 2500 “ “ “ “ 3: 1000 – 1500 “ “ “ “ 4: 500 – 1000 “ “ “ “ 5: < 500 “ “ “ “
Urbanity(sqkm)	0	The number of addresses per square kilometer
Industrial	0	0: not classified as an Industrial neighborhood 1: classified as an Industrial neighborhood
Neighborhood	0	The name of the neighborhood
Traveltime	0	The travel time to complete an order

Appendix C – Neighborhoods sorted on frequency

Neighborhood	Frequency	Inhabitants
Hurk	231	70
Blixembosch-Oost	119	7300
Binnenstad	107	3810
Zwaanstraat	104	595
Tempel	95	5095
Prinsejagt	84	4695
Flight Forum	81	0
Villapark	72	2075
Genderbeemd	68	3640
Grasrijk	65	5835
Woenselse Heide	64	5165
Woensel-West	62	3780
Veldhoven	60	5395
TU-terrein	59	810
Generalenbuurt	57	5415
Irisbuurt	56	2255
Hanevoet	55	3680
Het Ven	55	4045
Achtse Barrier-Gunterslaer	54	3735
Hemelrijken	51	3765
Strijp S	51	1665
Cobbeek en Centrum	49	4075
Achtse Barrier-Spaaihoef	47	4515
Vaartbroek	47	5225
Eliasterrein, Vonderkwartier	46	3175
Schrijversbuurt	45	3540
Schoot	45	2965
Kronehoef	44	4105
't Hofke	44	3470
Tuindorp	44	2925
Lievendaal	44	3150
Oude Gracht-Oost	43	1320
Heesterakker	42	2680
Oude Gracht-West	42	2835
Kerkdorp Acht	40	3490
Eikenburg	39	1505
Philipsdorp	38	3115
Doornakkers-West	37	3490
Meerveldhoven	36	2365
Lakerlopen	36	3290

Neighborhood	Frequency	Inhabitants
Burghplan	36	3050
Achtse Barrier-Hoeven	35	4005
Muschberg, Geestenberg	34	3980
Jagershoef	34	3575
Gildebuurt	34	1655
Drents Dorp	34	2385
Tongelresche Akkers	34	1235
Gerardusplein	33	3390
Barrier	33	2140
Genneperzijde	33	1380
Kerstroosplein	33	1870
Gijzenrooi	32	1830
Bennekel-Oost	32	3375
Eckart	31	4300
Kruidenbuurt	31	2970
Puttense Dreef	31	1240
't Hool	30	2240
Blaarthem	28	2445
Roosten	28	720
Ooievaarsnest	27	890
Genderdal	27	2935
Bergen	26	2620
Rapenland	26	2335
Bennekel-West, Gagelbosch	26	3400
Vlokhoven	26	3530
Mensfort	25	3065
Blixembosch-West	25	2095
Zeelst	23	5375
Oude Toren	23	1645
Luytelaer	23	945
Limbeek-Noord	23	2385
Hagenkamp	22	1190
Doornakkers-Oost	22	2870
Aalst	22	3640
Schouwbroek	22	1535
Driehoeksbos	21	975
Engelsbergen	21	640
Witte Dame	21	2030
Sintenbuurt	21	1800
Waterrijk	20	1695
D'Ekker	20	4080
Rapelenburg	19	865

Neighborhood	Frequency	Inhabitants
De Kelen	19	4080
Zandrijk	19	2975
Rochusbuurt	18	1775
Fellenoord	17	170
't Look	17	2735
Industrieterrein Ekkersrijt	16	30
Koudenhoven	15	500
Oude Spoorbaan	14	2060
Woenselse Watermolen	14	1345
Elzent-Noord	14	1080
Beemden	14	0
Schuttersbosch	14	590
Nieuwe Erven	13	1115
Joriskwartier	13	1270
Karpen	13	450
Tivoli	12	1430
Mispelhoef	12	25
Poeijers	11	0
Limbeek-Zuid	10	1410
Bloemenplein	10	1245
Nuenen-Noord	9	5470
Esp	9	5
Hondsheuvels	9	255
Breeven	8	25
Zonderwijk	8	3465
Heikant-West	8	3910
Winkelcentrum	8	655
Looiakkers	8	575
De Polders	8	2820
Elzent-Zuid	7	290
Park Forum	7	20
Vredeoord	6	490
Kapelbeemd	6	105
Eeneind	4	730
Verspr.h. Scherpenering en Landsaard	4	800
Sportpark Aalsterweg	4	15
Oerle	4	2745
Nuenen-Zuid	4	7095
Eckartdal	4	290
Bosrijk	3	415
Mierlo	3	9555
Riel	3	125

Neighborhood	Frequency	Inhabitants
Verspr.h. ten zuiden van de E3-weg	3	275
Nuenen-Oost	3	6145
BeA2	3	30
Castiliëlaan	2	65
Ekenrooi	2	4005
Verspreide huizen Zittard	2	275
Bokt	2	125
Meerbos	2	45
Verspreide huizen Son	2	1410
Heikant-Oost	2	2585
Wielewaal	2	90
Herdgang	1	10
Zesgehuchten	1	3470
De Gentiaan	1	4395
Waalre	1	6470
Urkhoven	1	165
Heivelden	1	3895