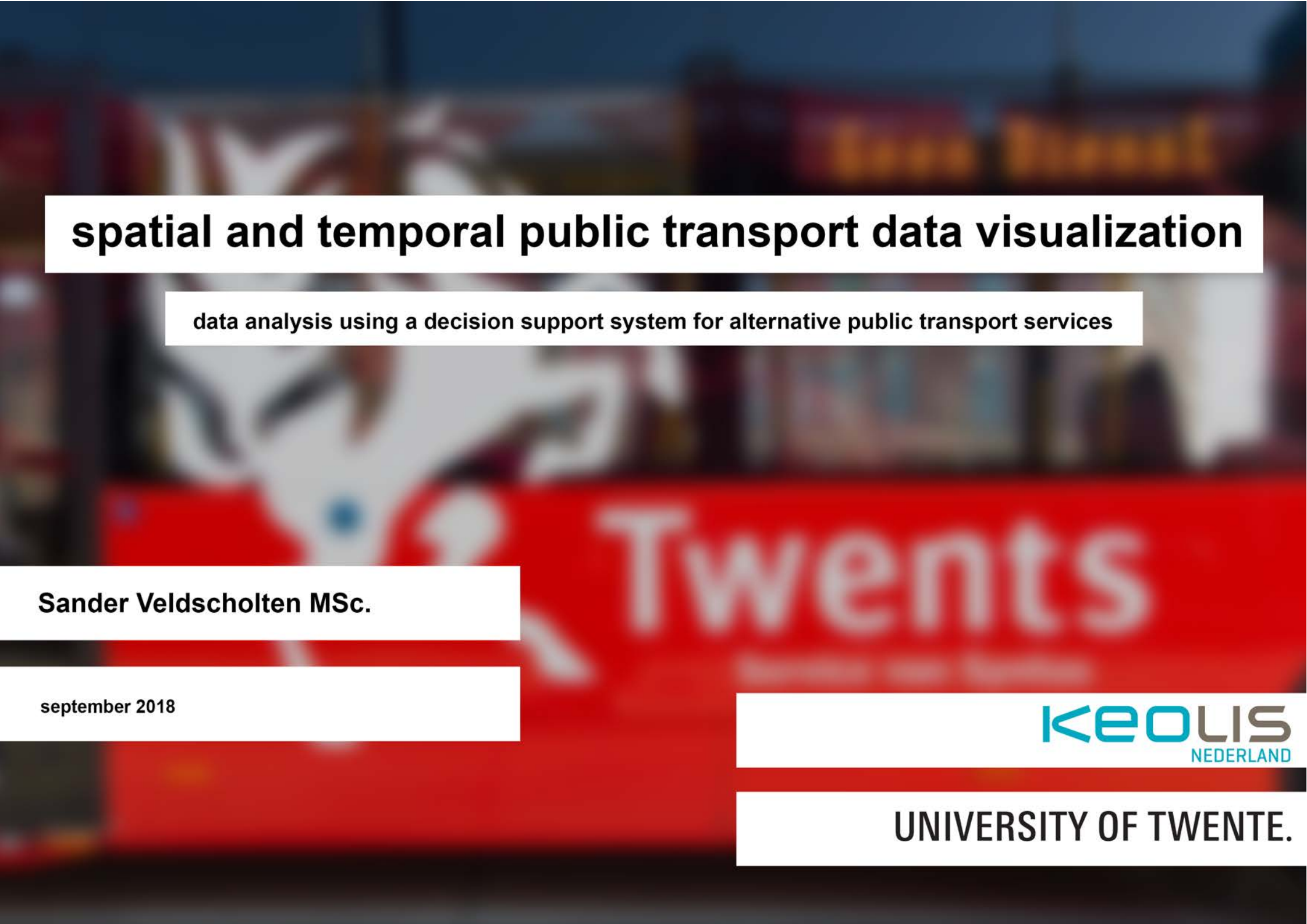


# spatial and temporal public transport data visualization

## data analysis using a decision support system for alternative public transport services

Sander Veldscholten MSc.

september 2018

KEOLIS NEDERLAND

UNIVERSITY OF TWENTE.

# Spatial and temporal public transport data visualization

Data analysis using a decision support system for alternative public transport services

PDEng thesis
Sander Veldscholten MSc.

University of Twente
Faculty of Engineering Technology (CTW)
Centre for Transport Studies (CTS)

PDEng assignment
September 2018

Supervisors:
Prof. dr. ing. K.T. Geurs
Dr. K. Gkiotsalitis

External supervisor (Keolis Netherlands):
R.R.M. Oude Elberink

# Preface

## Learning

After graduating in 2009 I knew I wasn't ready with studying. At that moment I did want to bring my studies in practice. I also decided that if I got the chance to go back to class I would grab it to broaden my knowledge. I'm very grateful for the chance that was given to me by Keolis to fulfil this promise to myself by sponsoring a two year post master PDEng programme at the University of Twente. Although I had my doubts starting a technical study with my acadamic social background, I quickly felt and knew I wasn't out of place at the centre for transport studies. The knowledge I obtained professionally in the field of public transport combined with my personal interests in programming and data were combined in this study.

The most valuable knowledge during the two year PDEng programme was for me to realize that there isn't an unbridgeable divide between technical and social educated professionals. The difference between them is just the way they perceive a problem. In my opinion the difference is mostly in the approach, a social scientist in general has a top down world view, whereas an engineer approaches a problem far more bottom up.

It took me quite some struggles, a few tough classes, a lot of dedication and self-motivation to get there, but I am proud to be able to call myself an engineer now next to being a dedicated social scientist.

## Aknowledgements

I would like to thank everybody involved in the completion of my PDEng programme. As a first I would like to thank my supervisor prof. dr. ing. Karst Geurs for taking time in his busy schedule to steer me in the right direction. Thanks Kostas for taking the time to help me in your very busy first months, I enjoyed giving you a tour into the world of Dutch public transport. And of course a lot of thanks to my roommates in the PhD-room, who are great stress relievers and place things exactly in the right perspective.

At Keolis Steven has been a great help in getting to know the particularities in the PT data. The colleagues right next to me, Rob, Victor, André, Simon and Berry, were of great importance to keep it fun to work on this project and their practical insight into the world of public transport helped me to focus my research. As a last I would like to thank Robert for supporting me the past few years and going out of his way to make this PDEng possible for me.

I also would like to thank Gertrud from the province of Overijssel for making the data on regiotaxi available for research.

On the personal level I would like to thank Liese for helping and supporting me outside of the working hours. It couldn't have always been easy, as it was at least for me, sometimes hard to separate work and home.

Also the people not mentioned here by name because you have temporarily slipped my mind, thank you for supporting me and taking the time to read this!

# Abstract


## Background

The field of public transportation is transforming. Whereas in the past public transport was organized top down, with services being offered and people being tempted to use these services, public transport companies are transforming into a more bottom up service provider where services are being offered which suit the needs of the potential customer. This change is driven by technological advancement and possible with the ever growing availability of data.

## Research

The objective of this study is to provide Keolis with an easy to use system which can be used to gain more insight in travel patterns of people using public transport in the Twente region in order to be able to offer services more tailored to the wishes of the customer. This is done by first inventorying and judging available and for this project useful data sources. Using the Knowledge Discovery in Databases (KDD) method as a guideline, the data was cleaned, transformed and analysed. Distance decay functions were generated for different modalities based on the national travel survey OViN and for public transport also on data from the OV-Chipcard, the national public transport payment card. Data from the regiotaxi service, an alternative form of public transport, was analysed as well as the users from this service are an interesting new source of users for public transport.

## Design

Next to doing research, the most important part of a PDEng programme is the design part in which a tangible product is to be delivered. A decision support system was built to visualize temporal and spatial travel patterns of public transport users.

Using the design cycle methodology a web based tool was designed in which it is possible to show factual travel relations and demographical data based on parameters which can be easily altered by the user. The maps generated can be used as a complementary source in proposals for changes in the service level or the creation of new or alternative public transport services.

## Case studies

To show the potential of the tool three case studies have been worked out in which the DSS was used to answer different practical questions. These case studies are based on concrete questions from within different parts of the company.

A study involving the municipality of Rijssen was done to see if regular public transport would compete with the new TwentsFlex service. Based on the results from the maps generated by the tool there is no competition between the two. The potential of the newly introduced neighbourhood bus in the municipality of Borne was analysed. This research is based on OVCK data and regiotaxi data. In the analysis it became clear (semi-) public transport in the municipality of Borne is mostly used for travels to outside of the municipality, which is probably because the distances within Borne are easily done by foot or bike. The potential for a neighbourhood bus therefor is deemed low.

The last case study was looking into the potential for a morning bus connection between Denekamp and Almelo to offer a more direct route for students traveling to school each day. Based on the data available it looks as if there is no potential at all between the two locations, which is completely contrary to intuition and signals from bus drivers. As the tool is to be used as a complementary source, the problem was looked into deeper. The conclusion was drawn that the data available at the moment in the DSS is not sufficient to answer this question using only the tool.

Using the tool, which is still in prototype or early alpha stage, already interesting and surprising conclusions can be drawn for concrete questions.

## Conclusion

Using only data which is available to Keolis for free, by using internal OVCK data, partner data from the regiotaxi service provided by the province of Overijssel and data found freely available on the internet, it is possible for Keolis to gain enhanced insight into the travel patterns of current and potentially groups of users. The decision support system, currently in a prototype or alpha state, can be developed further and has a lot of potential and, as shown in the case studies, can be used in the process of answering different questions.

Future development can be done by adding (travel) data from other sources to the tool. On the research part it will be possible to extend the tool with a potential travel amount estimation, based on characteristics of regions using regression or machine learning techniques. An extension of the tool can also be done by adding other sorts of data like customer satisfaction data, turning it into a versatile spatial data visualizer which can be used in KPI monitoring as well.

# Table of Contents

# Lists

## Figures and tables



## Abbreviations

| | |
|---|---|
| DM | Data Mining |
| KDD | Knowledge Discovery in Databases |
| OD | Origin Destination |
| OVCK | OV-Chipkaart |
| PT | Public Transport |
| PTA | Public Transport Authority |
| RT | Regiotaxi |

## Definitions

| | |
|---|---|
| Trip | - time in / on one specific vehicle / modality<br>A bus transfer is counted as two or more trips |
| Journey | - a combination of trips from origin towards destination |

# Introduction

## Background and motivation

### Keolis Netherlands

The public transport company Keolis Netherlands has expressed the ambition to transform from a classical public transport company, which is more or less organized top-down, into a provider of mobility services which is far more client, or bottom-up oriented. For this transformation to be successful, it is crucial to have a thorough insight in the publics' needs for travel. For Keolis therefor it is very beneficial to be able to estimate the directions and timeslots in which people in a distinct area travel. These temporal and geographically bounded 'corridors' can be used to predict when, from and where to there is a potential need for travel.

If there is enough potential mass on a corridor to make it financially interesting for the public transport company, the information obtained can be used as input for a proposal to offer a new service which meets the demands of the public. Depending on the geographical and demographical characteristics of the area, a specific type of (alternative) public transport can be proposed. Alternative services consist of, but are not limited to, for example a rush-hour service, a neighbourhood bus, a shared taxi system, bike sharing etc. In short alternative public transport includes all services which could be provided by Keolis which are not a regular bus or train on a time table.

For Keolis, different motivations for the transformation towards a mobility services provider and thus research in travel behaviour can be discerned;

### Changes during concession period

By law ("Wp 2000," 2000, 6 juli) Public transport (PT) in the Netherlands is being executed by a public tender. Concessions are being put on market by the responsible government agencies for periods of about ten years. Companies interested in offering their services to the area which is up for tender, can prepare their bid by the guidelines the responsible government publishes. The company which offers the best bid, which is a combination of being the cheapest for the tendering party or the one which offers the highest level of service, has the exclusive right to offer public transport in the area during the period the concession contract is in effect. As preparing a bid and implementing the service takes quite a lot of time and resources for a public transport provider, there could be up to twelve or even more years between the moment of starting the preparations for a bid and the end of the concession period.

The request for tenders for the concession Twente for example, came to market in 2010 and will only end in December 2023, making this a long term commitment lasting about 13 and a half years, not even including possible extensions of the contract and the preparation time for the government agency preceding the publication of the tender, which can also take years.

As concession periods are long, a lot influencing the service level can and will change in the time the contract is in place. In periods spanning over a decade, urban planning becomes a factor to take into account as new neighbourhoods, shopping centres and industrial areas are built or relocated for example which impact the need and direction for travel. The economic business cycles, or also called Juglar cycles, which as a rule of thumb span seven to eleven years (Korotayev & Tsirel, 2010), have an influence on the demand for transportation as in optimistic times the demand for transport is higher than it is during a recession. Technological advancement impacts travellers expectations. In addition, fuel and ticket prices can have dramatic effects on profitability. Policies by public authorities can also potentially have a big influence on travel behaviour, as PT is paid for by public grants for a big part in direct subsidies and government paid student subscriptions.

Due to a variety of unforeseen reasons, the ten year concession ‘Twents’ which was awarded by Regio Twente in May 2011 and which Syntus (now Keolis Netherlands) started the exploitation for in December 2013, was making quite some structural losses already in 2015. To tackle this problem Syntus started with re-evaluating the PT service level and demographic structure of the Twente area by using the Neolis method. This research program developed by Syntus’ main shareholder, the French PT company Keolis, was developed to analyse if supply and demand of public transport in an area is in balance and if bus services were offered from and to the right places. The use of a research method which was tested and used in an international context (as it was in use in Stockholm for example) helped in the negotiations with the provincial government to open up the concession contract and to be able to make changes in the service level offered, to increase profitability.

In hindsight, the Neolis program wasn’t a good fit for the questions Syntus had for the Twente region. Nevertheless, doing this research helped in opening up negotiations with the province. These negotiations and the resulting changes in the service level turned the financial tide for this concession in 2017 already. Also lessons learned from this research program can be used to develop a research method which answers the questions at hand in a better and more effective way for the, in an international context, rather unique Dutch public transport situation as there is a very high acceptance of the bicycle as a transport modality.

## Information and expectation

When did the always connected smartphone become mainstream? This question is extremely relevant to understand the changes in processes in the PT world.

The beginning of the smart phone era started some ten years ago with the first version of Android and the iPhone 3G, which could be seen as the first mainstream smartphone from Apple, which were both released in 2008. This means the personal information age’s “early majority phase” started not even a decade ago if you take

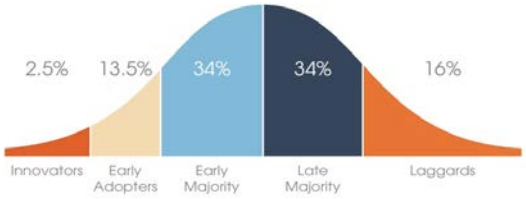


*Figure 1: Rogers' bell curve of technological adoption (Rogers, 2003)*

into account the diffusion of innovation bell curve (Figure 1) by Rogers (2003). Currently about 89% of Dutch inhabitants, over the age of 12, have a smartphone (Telecompaper, 2017), whereas the penetration in the lowest age group (12-19) is even 97%, compared to 73% in the highest age group (65-80).

The smartphone caused a revolution in the availability of data and information. With this availability of information, opportunities for new ways of collecting and presenting information arose. New companies came up with revolutionary, or not so revolutionary but well executed, ideas and with this the publics’ expectations of classical services changed.

In about ten years’ time we have gone from a printed bus booklet, to a real time updating route planner application in nearly everybody’s pocket. This planner takes into account delays and recalculates a more efficient trip on the fly if for example a bus malfunctions. Currently tests are being executed with full service mobile travel planners which take it even a step further. These new applications take into account personal preferences, weather conditions, travel time and costs for (combined) travels by cars, public transport or bike, giving the user the most cost or time effective journey available.

Technological progression offers a lot of new (business) possibilities. Because of these possibilities expectations of people, the users of these inventions, change as

well. In public transport for example, only ten years ago it was still perfectly normal to own a physical timetable booklet if you used the bus or train quite often. Nowadays it is in some concession areas not even mandatory to physically print these booklets as patrons are used to and expect to have real time updates on their phone on bus and train punctuality in minutes and preferably seconds. On social media, comments about busses arriving only one or two minutes later than planned are quite common nowadays. It is even possible to follow buses and trains on a real time map online. For cars there is dynamic navigation available, recalculating a new route in real time if it detects a traffic jam has originated on the route. In the near future we will have apps which integrate multiple mobility services, and dynamically suggests the mobility mix most efficient (cost and/or time) for a specific journey. This new access to information changes the way people can and probably will travel.

Due to the availability of information, travel becomes more and more tailored to the individual needs of the traveller (bottom-up), whereas in a classical public transport company the main focus is just the other way around, on collective travel (top-down). With a better insight in actual travel patterns, a public transport company can implement alternative public transport modalities or services to tempt people to use a more personalized type of public transport.

Competing companies are also recognizing the value of data and information. In an interview (Clahsen, 2017) the CEO of Connexxion stresses the need for innovation by using data and business intelligence in public transport to be able to survive as a PT company in the long run.

### The future of transportation

With the technological progress discussed in the previous paragraph, new companies come into existence. These companies already have or probably will have a disrupting effect on mobility as we know it. Companies like Uber, BlaBlaCar and Lyft, with which car sharing or carpooling has been made easy and even ‘sexy’, something which government and business campaigns didn’t pull off (Steenbrugge & Dedecker, 2015).

Another disruptive company is Tesla motors, which is, at least in general media, a front runner in self-driving cars. Self-driving cars could make taxi trips potentially quite a lot cheaper, as the biggest cost component (about 66%) in a taxi is the driver (van Beijeren & Dasburg-Tromp, 2010). Autonomous cars have the potential to dramatically reduce traffic jams by more efficient driving and thus increasing the network capacity. According to a study by Tientrakool (2011) the capacity of highways where all cars are self-driving can increase up to 273%. With statistics like these, self-driving cars (potentially) reduce traffic jams to a minimum and can quite possibly compete with regular public transport on price, both these properties decrease the demand for classical public transport.

Again, these new initiatives focus on the individual travel needs of the public. For a public transport company to survive on the long run, it is important to act on this trend and explore how these trends can be used as opportunities for the future of the company.

### Alternative Public Transport

Public transport in the form of twelve meter long buses on a fixed timetable probably isn’t what people expect any more as a regular form of public transport in the near future. As there is more and more information available to make travels personal, people will also expect more personalized means of transportation. A trip should start from their home, and only end until they have visited their destination and are home again. There are several ideas in the form of services or transport modalities, which a public transport company could use to extend their current service and to increase the service level for the customer. Ranging from shared bikes and neighbourhood shared cars to flexible bus services and self-driving shuttle buses. A completely other

way of serving travellers in the future could be a travel suggestion application which helps in choosing the most efficient means of transportation for a certain activity.

Different concepts and ideas for alternative public transport modes are available within Keolis International. Quite some experimentation is going on at the moment within the international Keolis-group on different types of alternative public transport services. Ranging from self-driving buses in Las Vegas to (Keo)bike sharing in the Netherlands.

## Company

Keolis Nederland is part of the French Keolis Group. Keolis Nederland started in 1999 as Syntus (*SYN*thesis between *T*rain and b*US*) in the Dutch region of the Achterhoek implementing a new concept in the public transport, the so called 'fish bone model' (visgraatmodel) in which the regional train time table was integrated into the bus planning. This concept was very successful in revitalizing the public transport in this rural area.

Now Keolis Nederland is a company which in the Netherlands provides public transport in four bus concessions, one combined bus - rail concession and one dedicated rail concession. With 2200 employees, 25 trains, 700 buses and a revenue of about €230 (with an ambition to grow to €300) million a year, Keolis now is a big player in the Dutch public transport market.

### KPI's

Key performance indicators used to determine the success of a service mainly depend on the contractual agreement with the public transport authority (PTA). Roughly 50% of the income in public transport are government funded subsidies. The other half are, in case of a revenue contract, direct income from passengers. Based on indicators in the contract a bonus-malus payment is in effect. The main indicators used in the concession Twente are punctuality, customer satisfaction and growth in patronage.

### Current state

At the moment alternative public transport for PT companies is mostly seen as a means of decreasing costs. If a regular bus is not profitable anymore due to lagging patronage, a solution is being sought which decreases costs for the provider but still gives a reasonable level of service to the low amount of current passengers. In that sense alternative public transport is currently perceived as something negative as it replaces a better but, for the provider, more expensive alternative.

At the moment costly ad hoc research has to be done every time a change in the service level is proposed. With a system which is being developed in this study much of the data which a proposal can be based on is already available, saving thousands of euros in hiring external research companies doing ad hoc research.

### Gap in knowledge

As explained in the sections before, Keolis has expressed the intention to transform from a public transport company into a provider of mobility. To be able to make this transformation a new kind of knowledge is required in the organization. Where in the current situation knowledge about infrastructure and effective scheduling is of utmost importance to efficiently run the operation, as a provider of a mobility service you want to be able to predict where and when people need your service to be able to offer it at the right time in the right place and in the right amount.

At the moment when an alternative mode of transport is being proposed it is mostly as a replacement for a bus line which isn't profitable to run anymore. Based on the amount of passengers in the last few months on this line an alternative mode is proposed. By using OVCK check-in data combined with a rough estimation of single ticket sales on the bus, an estimate for the last few months is made for the amount

of passengers traveling the line which is up for cancellation. Next to this estimation based on available data, most of the times a manual count is done for a few days by a research company to be able to check if the data is correct. This last part of the research is quite expensive, as multiple people are needed for quite some hours to be able to do a full count on one line which travels in two directions multiple times per hour. It should be possible to skip this last (manual) research by more advanced data analysis to bridge the gap between data and reality. This research project aims to offer an extra means of information which can be used to bridge the gap in knowledge when lines are up for cancellation. This could help in reducing the costs for a manual count.

The biggest gap in knowledge for Keolis actually lies in the unexplored possibilities. How can corridors be found where it is potentially most lucrative to start a new service based on the travel patterns by inhabitants? There currently is no clear method on expanding the amount of services within an existing concession. This research aims to give an overview of the area of Twente on travel patterns of the inhabitants in this area. By combining data from different sources the potential amount of travels will be estimated between postal code zones. This information can in turn be used in proposals for new public transport services in the area. In the conclusion three case studies can be found in which a start has been made on explaining travel patterns in an area based on the data available.

In short, this research aims to tighten the gap in knowledge at Keolis on public transport travel patterns to be able to offer new services on corridors which could be profitable due to a potential high demand. This can be extended into understanding future trends in public transport use to be able to make the transition to a more bottom-up approach of public transport.

## Outline of PDEng thesis

This thesis has started with an introduction into the problem which it is trying to solve. In the next sections first the objectives will be described; this includes a description of the issue as well as a description of the objective of the PDEng programme. This is followed by a programme of requirements in which is described what the conditions are the final product needs to be built on. The next chapter will be a literature and data review which will describe methods used from a theoretical perspective and gives an overview on the data which has been considered and which data was used in this research. Followed by a chapter about design methodology which will be about how the research theoretically will be done. Needless to say this chapter is followed by a chapter about the actual development. This research will end with case studies, a conclusion and recommendation on future possible developments which will extend the current project.

# Objectives

## Description of the design issue

The objective of this study is to provide Keolis with a system which can be used to gain more insight in travel patterns of people using public transport in the Twente region in order to be able to offer services more tailored to the wishes of the customer. The information in this tool can be used in proposals for new services or changes in existing ones, this is shown in the case studies chapter of this report. At the moment there is no structural process or tool available within the company to be able to easily visualize spatio-temporal public transport travel patterns. When it is necessary to access this information in order to make changes to the PT network, mostly ad-hoc solutions are used. The current process of retrieving information on travel behaviour or patterns can best be described with a starting point of 'gut-feeling' followed by ad hoc data requests at the IT-department on current patronage and if necessary an (expensive) passenger count by a research company.

All considered, the issue to be solved with this research is the absence of a structural means of clear insight for Keolis in the travel patterns of the people living in the areas serviced by Keolis Netherlands.

## Objectives of the design project

- Travel and demographical data identification, appraisal and preparation
- Data mining in order to create OD-matrices
- Designing an easy to use tool to display travel patterns in Twente

# Programme of requirements

Part of a design assignment is to look into the requirements the design has to meet in the end. Using the format in the PDEng thesis template, different requirements were looked into. The advantage of creating a programme of requirements on beforehand is that when the product is in development it will be clear what the focus should be.

## Safety/Risks

- The tool should not be able to change data, as this will be a risk in data integrity.
- Privacy issues according to the GDPR should be dealt with.

## Reliability

- The tool developed is an indicator and only one of the tools at disposal for Keolis to base advices concerning new means of or changes in existing services on. This means the results the tool present don't have to be 100% reliable.

## Maintenance

- The tool should be low maintenance as knowledge and budget will not be available anymore to do maintenance after the completion of this project.

## Finances/Costs

- As cheap as possible, in house data or sources which are available for free as there currently is no extra budget to continue developping the tool.

## Legal requirements

- Results should not be possible to lead back to one individual (GDPR)
- Internal data should stay internal

## Environmental/Sustainability

No requirements could be formulated on this subject.

## Social impact

- The tool, when properly introduced can have an impact on the workflow of people involved in changes in service level. It should be non-invasive, as then it would probably not be used.
- Easy to use is key. A person involved in making the first plans into changes in service level is in practice most of the times not a specialized technical person. A few clicks and the result should be visible. Long waiting times are not acceptable.

## Recyclability/Disposability

- Code can be reused or recycled for use in other applications. Source code has to be readable.

## User

- The main user will be non-technical, tool has to be intuitive and easy to use.

# Literature review

In this part of the report methods and theories which were used in this project and the data analyses will be elaborated on.

## Method: Design cycle

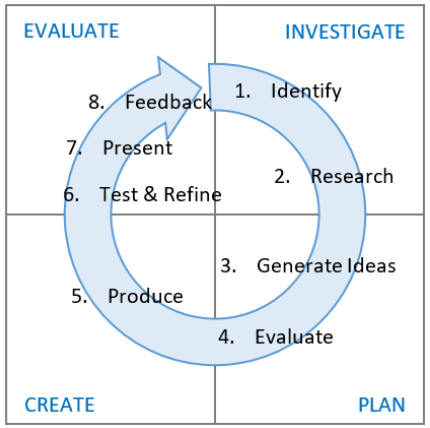


*Figure 2: Design cycle representation*

In the education part of the PDEng programme the design cycle has been introduced in multiple courses as a means to structure a design process. This theory of approaching and managing a (technical) project has many different (sub-)methods and tools for different phases in the cycle which have been developed by a lot of different researchers. The baseline actually is more or less the same in all approaches though, they vary in complexity like how much information or steps are shown and in specific wording. The cycle as shown in Figure 2 is a simple graphical representation of the design cycle which is being used as a framework to structure this project.

The design cycle will be explained more thoroughly in the next chapter; “Design Methodology”

## Method: Knowledge Discovery in Databases

KDD, an acronym for Knowledge Discovery in Databases, refers in short to the process of finding information or knowledge in data sets. This theory has been used as a guideline in this project to be able to transform data into information. Although KDD is a quite commonly used terminology in the field of data science, there is quite some confusion about what it actually encompasses when literature from different authors is being compared. The most confusion seems to be about the relationship between data mining and KDD as for example there are authors describing data mining as a part of KDD;

> *“KDD refers to the overall process of discovering useful knowledge from data, and data mining refers to a particular step in this process. Data mining is the application of specific algorithms for extracting patterns from data.”*
> *(Fayyad, Piatetsky-Shapiro, & Smyth, 1996)*

Also sometimes authors use the terms KDD and data mining as a synonym, and also change the meaning of the acronym a bit in the process;

> *“Data mining, also popularly referred to as knowledge discovery from data (KDD), is the automated or convenient extraction of patterns representing knowledge implicitly stored or captured in large databases, data warehouses, the Web, other massive information repositories or data streams.”*
> *(Han, Pei, & Kamber, 2011)*

And sources can be found where KDD is being described as part of the data mining process;

> *Knowledge discovery in databases (KDD) is the process of discovering useful knowledge from a collection of data. This widely used data mining technique is a process that includes data preparation and selection, data cleansing, incorporating*

*prior knowledge on data sets and interpreting accurate solutions from the observed results.*
*(Menken, 2013)*

For the current study the first description in this chapter, the one formulated by Fayyad et al. (1996) is used. KDD is a process of which "data mining is the application of specific algorithms for extracting patterns from data" (ibid.).

Knowledge Discovery in Databases in the study by Fayyad et al. (1996) is transformed into a five step process;

1. Data Selection
2. Data Pre-processing
3. Data Transformation
4. Data Mining
5. Data Interpretation / - Evaluation

This five step process is used as a guideline in the methods chapter to describe the steps taken in this study for creating the decision support system based on data from different sources. In theory these steps follow one another. In practice it can be more efficient to switch some steps around and for example first do a data transformation before going to the pre-processing phase.

### Data Selection

The data selection phase can be described best with the question: Which data is used? This depends on the availability of relevant data and on the goal and budget of the project. As selection of data is an integral part of the method used in this project, the description of the data considered for this project will be described in the development phase chapter of this report.

### Data Pre-processing

Data pre-processing is mostly about cleaning the selected data. Real world data most of the time is incomplete, noisy and inconsistent (Malley, Ramazzotti, & Wu, 2016). In the pre-processing phase, missing values are addressed by deciding what to do with them. Accepting them and working around or by using mean values to fill the gaps for example.

It is possible to use regression or machine learning techniques in this phase to clean the data or to fill in missing values by using (un)supervised methods or regression algorithms.

### Data Transformation

Data transformation, the middle part of ETL (Extract, Transform and Load) is basically the transformation of data from one structure / source or format into another. When data from different sources need to be combined, they are rarely in a format which is usable for the tools used in the next step; data mining. Therefor data needs to be converted into formats which can be used for integration and analysis.

Although some research is being done to automate conversions using learning algorithms for normalization purposes (Wu, Sekely, & Knoblock, 2012), in practice this method is not accurate enough to be usable. The transformation process thus is to be done by using ETL-tools, or by writing transformation scripts. An advantage by doing the transformation by hand instead of using algorithms is that anomalies and potential connections between parts of the data can be spotted earlier.

### Data Mining

In this phase data is being analysed by using techniques and algorithms. Data mining techniques fall in different classes. Classes of techniques interesting for this project are;

- Association rule learning, also known as market basket analysis,
- Regression, to find a function which predicts relations among data(sets),
- Summarization, to present the data in a compact way using reporting and visualization techniques.

In this project the summarization part of the data mining phase is used mostly, as the objective of the project is to present the data in a way non-data analists can work with it. This means it has to be clear, easy to use and free of hard to explain methods.

### Data Interpretation / - Evaluation

Last in the KDD process is the interpretation and evaluation phase. In this project a few cases will be used to explain some phenomena which become apparent when using the data in the tool developed. As data is always a representation of the real world and not the actual world it is necessary to indicate what the results actually mean and what the limitations are of the information shown.

## Theory: Mode choice factors

What factors determine the use of modality and thus public transport and thus are interesting factors to include in the development of the decision support system. The meta-study by Hollevoet et al. (2011) have identified and structured 23 important determinants which influence modal choice. Hollevoet et al. (2011) have split up the interacting determinants into four pillars (Table 1), each representing a few different important determining factors

These determinants are the most important factors on which a choice is made by a person for a specific type of modality when a journey is being made. For example if the determinant distance grows, walking and cycling will decline, whereas the use of car, train and plane will increase with distance. The determinants also influence each other, density, or the amount of people living in a confined space for example influences the availability of PT (proximity to infrastructure) and the severity of the rush hours (travel time).

| SPATIAL | SOCIAL-DEMOGRAPHIC | TRAVEL MODE & JOURNEY | PSYCHOLOGICAL |
|---|---|---|---|
| Density | Gender | Distance | Habits |
| Proximity to infrastructure | Age | Travel time | Experiences |
| Parking | Employment | Travel cost | Perceptions |
| Frequency of PT | Income | Trip chaining | |
| Diversity | Lifestyle | Departure time | |
| Interchange | Education | Travel motive | |
| | Household size | | |
| | Car availability | | |

*Table 1: Determing factors in modality choice Hollevoet, De Witte, and Macharis (2011)*

There are methods to quantify the influence of each or in most cases some of these determinants to predict the modality choice in a certain situation. In case of public transport planning this can be used to determine where and when there is potential for a PT service. At Keolis the research program Neolis is an example in which indicators for proximity, frequency of PT, employment and car availability are combined to determine the gap between demand and supply in public transport.

## Theory: Distance decay function

In essence a distance decay function shows the willingness of people to travel a certain distance or period of time. In practice this means the bigger the distance between two locations the less people are willing to travel between them. This makes this function a practical application of Tobler's first law of geography; "everything is related to everything else, but near things are more related than distant things." (Tobler, 1970)

With this function it is possible to show spatial interaction between different locations based on distance. When enough data is available specific distance decay functions can be constructed on regional, local or time based level. One premise which can be tested with a distance decay function could be that people living in more rural areas tend to accept a longer travel time or distance compared to urban inhabitants. This

is based on the idea that services are wider spread in rural compared to urban areas. Next to difference in location or time of day also difference in willingness to travel a certain distance can be differentiated on modality.

To create the distance decay function the amount of trips made per time interval are summed up. This data is transformed into 1 - cumulative percentage. When plotted, this already is an observed distance decay curve. Using this data, a regression can be performed which approximates the observed data and can be used in calculations.

## Theory: Smart card data

As smart card data from the OV-Chipkaart (OVCK) is the main data source in this study, previous research involving data analysis on smart card travel data was looked into. Although the primary use of smart cards in public transport is the collection of fares, a lot more can and is done with the data which is generated in the process.

Research in the field of smart cards can be roughly divided into four different categories (Pelletier, Trépanier, & Morency, 2011), hardware, implementation, data use and commercialization. For a study involving travel behaviour like the study at hand, previous research done in the data use part is most interesting. This part in the literature study by Pelletier (2011) is divided into three subcategories; strategic level (passenger

behaviour), tactical level (service adjustment) and operational level (performance) studies. The more recent literature study by Li, Sun, Jing, and Yang (2018) on destination estimation using PT smart card data gives an indication on the amount of work done on the strategic and tactical level. Over 200 unique papers were identified which are related to data analysis in PT.

As a lot of smart card systems implemented in public transport systems around the world don't require the user to check out at alighting. Therefor a lot of research is done into the field of destination estimation (Kurauchi & Schmöcker, 2017). As it is necessary for the user of the OVCK to use the smart card at boarding and alighting the exact O-D is already known. Therefor it is possible to take the data analysis step further and analyse for example differences in patron behaviour in space and time (Alsger, Mesbah, Ferreira, & Safi, 2015). This data in turn can be used to rearrange the network and schedules to better accommodate patron needs (Hofmann, Wilson, & White, 2009) or make a forecast based on historical data (Kurauchi & Schmöcker, 2017).

Raw data in the OVCK database is shown in a simplified manner in Table 2. The characteristics of the OVCK use combined with characteristics of the line (destinations of bus stops) and more data on Chip ID level, better predictions can be made on the characteristics, personal and travel pattern, of the user. These predictions in turn can be used to make the level of service better. As knowing the

| Chip ID | Check In StopID | Check Out StopID | Check In Time | Check Out Time | line | … | Ticket Type |
|---|---|---|---|---|---|---|---|
| **1001** | 35 | 488 | 2018-01-04 10:27 | 2018-01-04 10:52 | 9 | … | Regular |
| **1002** | 23 | 86 | 2018-01-04 8:01 | 2018-01-04 8:09 | 1 | … | Student |
| **1002** | 86 | 90 | 2018-01-04 8:17 | 2018-01-04 8:55 | 3 | … | Student |
| **1003** | 73 | 94 | 2018-01-04 7:20 | 2018-01-04 7:53 | 4 | … | Annual |
| **1003** | 94 | 73 | 2018-01-04 16:55 | 2018-01-04 17:27 | 4 | … | Annual |
| **…** | … | … | … | … | … | … | … |

*Table 2: OVCK database representation adapted from Kurauchi and Schmöcker (2017)*

needs of the customer makes it easier to sell them the right product or in this case (public) transport service.

**Side note on the use of OVCK data**

Using the data in Table 2 it is possible to construct the travel pattern of the fictional users;

- User 1001: Travelled only once on the day analysed
- User 1002: Made one trip with two connecting buses at station number 86
- User 1003: Travelled to station 94 in the morning and back to 73 in the evening

This travel pattern can be enriched by using a survey to assign value to properties in the database using regression or a machine learning algorithm. Using common sense (instead of a regression value based on a survey) on the fictional database above, the following assumptions could be done;

- User 1001: It is quite reasonable to assume this is a person not using PT that much as they travelled on a single ticket, only one way outside of the rush hour.
- User 1002: Probably a student travelling to school. Partying afterwards (as it is a Thursday) and traveling home after midnight.
- User 1003: A person travelling to and from their work as they have an annual ticket and travel only during rush.

# Design methodology / Design steps

This chapter has been cut up in four sections; Investigate, Plan, Create and Evaluation. These sections describe the methodology used in these different phases of the design project.

## Investigate

### Stakeholder-analysis

Below (Figure 3) the stakeholder analysis can be found. Stakeholders have been put in three different groups, directly involved, indirectly involved and possibly involved in the future. The work of Alexander (2005) on the taxonomy of stakeholders was used to create this schematic. By developing this analysis a better insight was gained in relevant people and institutions which in turn gave a better focus on where to put attention during the development process.

The most important stakeholders don't necessarily are the ones directly involved in the project. For example without data providers there would be no research possible at all. Without a mentor it would still be possible to finish the system. Graduation would be impossible without the mentor though, but that is a different stakeholder analysis.

Interesting stakeholders which require a bit more attention are for example the bus drivers which are being characterized as a threat agent. Changes in the field of public transportation nowadays are mostly not in favour of the bus drivers. New initiatives like neighbourhood buses or (kolibri/flex) taxi services are mainly driven by volunteers, and therefor a threat to job security for current bus drivers. A system which makes it easier for 'headquarters' to implement new public transport initiatives, which don't necessarily have to be new regular bus lines, can be perceived as a threat to current drivers. To manage this potential threat also the possibility to analyse travel patterns on current bus lines is stressed, next to the chances the tool offers for potential changes in routes which were proposed by (representatives of) the bus drivers themselves. Therefor also a case which was proposed by the works council (ondernemingsraad); a direct line in morning rush between Denekamp and Almelo, is added in the concluding chapter of this report.

| | Developer | Commercial manager | Manager alt PT | UT mentor | Board of dir. Keolis | IT Keolis | IT/revenue manager | External data providers | Bus drivers | Provincie Overijssel | Competitors | Travellers | Journalists | Tender Team Keolis | Keolis international |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Normal operator | x | | x | | | | x | | | | | | | x | x |
| Operational support | x | | | x | | x | x | | | | | | | | |
| Maintenance support | x | | | | | x | x | | | | | | | | |
| Functional Beneficiary | | x | | x | | | | | | x | | | | | x |
| Political beneficiary | x | x | | x | x | | | | | x | | | | | x |
| Financial beneficiary | | | | | x | | | | | | | | | | x |
| Negative stakeholder | | | | | | | | | x | | | x | | | |
| Threat agent | | | | | | | | | x | | x | x | x | | |
| Regulator | | | | | | | | | | x | | | | | |
| Champion/sponsor | | | | | x | | | | | | | | | | x |
| Developer | x | | | | | | | | | | | | | | |
| Supplier components | | | | | x | | | x | | x | | | | | |
| | Direct | | | | Indirect | | | | | | | | | Future | |

*Figure 3: Stakeholder analysis*

The possibility of cases like these have been stressed in contacts with the bus drivers representation.

The same can be stated for the travellers and journalists. Although changes in level of service are carefully done following strict procedures with the PTA and interest groups, there will always be someone who perceives negative consequences. Even in situations where a lot of positive effects can be quantified, it is possible that the story of one negative effect can overshadow the whole. By using a tool which uses data, the perception can be that the human factor is left out of the equation. This story can become powerful and be a threat to the system developed. News following the use of this tool therefor has to be managed by communication professionals.

### Socio-economic context

This project can have quite an impact in socio-economic context on the long term. The way public transport is organized at the moment is changing (Schmeink, 2018), also at Keolis. From top down organization where planning lines and more or less influencing people by marketing to make use of the PT services, towards a more bottom up approach where a means of transport is being offered for a trip the traveller wants (or needs) to undertake. This research project / design will play a role in this reorganization, as it will give information about the travel patterns of inhabitants on which new services can be offered. The introduction of new services based on the DSS, which is the end goal of this project, can lead to new jobs, but also make existing jobs change or even disappear. New jobs can be created in the form of bicycle maintenance, part time (small) bus drivers, IT-technicians for the development and maintenance of travel apps. Jobs which will change or could become obsolete are those of people conducting surveys, as more can be done with data which is already being collected. Also some bus routes could become obsolete, if this is the case less bus drivers could be needed which will be a threat to their employability. On the other hand it could also be that new routes will be introduced if chances for new connections are spotted using real travel data.

Newly implemented transport services, can also change the way people are traveling. Where people now are going to their destination by car, with a good offer (time and cost-wise) some of them can be persuaded to use a form of (semi-)public transport. Which in turn has an effect on traffic jams which in turn influence environmental pollution. Only a small amount of people need to be persuaded to not travel by car on their own, to make a huge difference in traffic jams “If 2,4 percent of people would carpool during rush hours, traffic jams will be 5-12% less” (van Wee, 2012).

On a smaller scale this research / design can change the work of people involved with implementing new forms of alternative mobility. With a method which is cheaper, faster and more accurate than the methods used now, their work could become easier. A tool, developed at a university can be good ammunition in negotiations with governments for example. It could also save money in the long run, as there is less need for hiring expensive consultancy firms to do large scale surveys. Money which in turn can be invested in a better performance of the mobility services.

Concluding; The creation of this DSS has an enabling effect on the change which is already going on in the way PT is organized. It will help in the transition of a top-down oriented organization into one which is more bottom-up. This in turn has an influence on the modalities people use to get around, which in turn has an influence on the economy and the environment.

## Plan

The problem at hand; the transformation process of a classical public transport company into a provider of mobility services, requires more insight on travel behaviour or patterns of the public. Possibilities to do so would be to, for example, develop a standardized method of doing periodical surveys, do panel discussions, or track people using a dedicated app on their phone and analysing the results.

With the still quite uncultivated area of travel data analysis and the interests of the researcher, the method chosen is to design a decision support system which uses (big) data sets to try and find geographical and temporal corridors in which it could be financially attractive to provide a form of new public transport.

### Objectives

The objective of this study is to provide Keolis with a system which can be used to gain more insight in travel patterns of people using public transport in the Twente region in order to be able to offer services more tailored to the wishes of the customer. The information in this tool can be used in proposals for new services or changes in existing ones. During development the focus area of the tool will be rural areas in Twente.

As there is a time constraint on this project, and it is not quite clear how much can be done in the time given there is a baseline objective which has to be finished within time and there are “bonus” objectives which will be developed if time allows them to be researched and designed properly.

*Baseline objectives*

- Identify, appraise and use different data sources on travel frequency and behaviour / patterns.
- Combine data sources to get enriched information on travel corridors using geographical and temporal parameters.
- Develop a tool which assists in the analysis of travel behaviour of pattern data for a pre-defined area.

*Bonus objectives*

- Update the tool with a function to find out which areas match certain criteria, so the tool can inform on locations where a certain type of alternative public transport could be successful.
- Organize an experiment based on the results of the analysis.
- Test the tool in a different region and update accordingly.
- Make it a complementary or maybe competing research method to Neolis, which is used at the moment within the Keolis-group to do similar research (internationalization).

### Influence stakeholders

As there are quite some direct and indirect stakeholders in this project, a strategy on how to handle these is of great importance. The approach to all stakeholders involved is an informal one with periodic meetings on milestones.

In practice this informal work style means working at the location which is most relevant at that time for the project interaction with the people involved during normal working hours is easy without having to plan formal meetings. During research the main focus will be at working at the university and keeping relations well at Keolis by working there at least one day a week right next to the people who are going to use the tool I’m developing. During development more input from the people at Keolis will be necessary so most time will be spent there.

## Create

### Data manipulation, enrichment and analysis

To actually built the decision support system data needs to be collected, analysed, manipulated, stored and enriched. The focus during development will be on documenting the principles used, as the end result is interesting, but the process is maybe even more important as the tool is more a proof of concept which later on could be integrated into the processes of the company. The DSS during development will be a (local) web application. By documenting the principles which are mostly SQL-queries used, it will be easier to integrate the tool in other applications later on. The documentation can be found in this report in the next chapter; “development phase”

## Evaluate

Test, present and feedback are the steps which are defined in the design cycle for the evaluation phase. Reading between the lines of the previous chapters, it becomes clear that evaluation is an integral part of the whole research and design of the DSS.

By using stakeholders' wishes as a base for the design of the DSS the evaluation process is one which is used continuously. By showing progress to the people involved they come up with comments, wishes and requirements which can be implemented during development.

# Development phase

## Knowledge Discovery in Databases

Knowledge Discovery in Databases (KDD) is a series of structured methods of turning (large) amounts of data into a coherent data source from which information can be gathered. More information about the theory of this method can be found in the theory section of this report. The sections below describe the concrete steps taken in this study from selecting to cleaning and transforming the data into the data warehouse used in the analysis. These steps consist of; data selection, pre-processing; transformation; mining and finishes with interpretation.

### Data Selection

This section describes the data sources and whether they have or not have been chosen for use in the development of the decision support system.

This research aims to be as cost effective as possible. The added value of a paid data source has to be significant to be considered for buying a license. In the section below all data sources which were considered for use in this project will be described along with the consideration whether to or not to use this data in the design of a data warehouse for the decision support system.

An overview of the data sources considered and the conclusion for use in this project;

- OVCK used
- Single ticket sales not used
- Regio Taxi used
- Demographics used
- Mezuro not used
- OViN used
- Geographical vector polygons used
- KNMI weather data not used
- Social media data not used
- Base map used
- GTFS time table data not used
- Bus stop data used
- Jobs used

#### *OVCK*

The OV-Chipkaart is the most used payment method in Dutch public transport. People check-in by placing an RFID-card on a reader when they enter the bus, or the station when traveling by train, and there has to be a check-out action when leaving or transferring a bus or station. These transactions are stored for billing purposes. It is also possible to use this data to construct origin destination matrices to gain insight in the travel patterns of public transport users as it is stored when and where on bus stop level the check-in and -out actions have taken place.

As transactions are stored for every check in and check out, these trips need to be combined into journeys when a person is hopping buses. When an OV-chipkaart is used to check in twice within 35 minutes this is counted as one journey. These 35 minutes are derived from the business rules in the OV-Chipkaart, as they also use this amount of time for a free transfer between buses (OV-Chipkaart, 2017).

It is also possible to travel by buying a ticket at the driver or by using the mobile app for example, therefor not all tips are being registered. There are methods to correct the numbers by using a multiplication factor, although internal research at Keolis Netherlands shows the methods used at the moment are not yet reliable enough. This means the OVCK data does not consist all trips made by passengers in the PT. The OVCK data is purely travel data from passengers who used the card as a means for payment.

In this study exports on trip data have been used from March 2017 and March 2018 in the concession of Twente to do the analyses and build the tool around. The analyses can easily be redone using a different timeframe or concession by exporting a different section from the data by changing parameters in the SQL-query which

downloads the data from the database. With a little more effort the tool can also be altered to query the full database and used semi-realtime once it is out of the prototype phase.

The months of concession data which were used during development were deemed to be a suitable representation of the data available. March is in the PT-world considered a month in which the new time tables have become routine, with relatively few traffic disturbances, vacations or other parameters which could be of influence on the data.

### *Single ticket sales*

Starting July 1st 2018 buses in several concession areas are ready for cashless transactions and it won't be possible to pay for a ticket by using cash anymore. This means all single ticket sales will be time and location logged starting July 2018. Thus in theory it will be possible to get even more complete PT-travel data. Although in practice it won't be possible to get this data within the timeframe of this research, as there is no export functionality available yet in the dedicated software to export the single ticket PIN-transaction data and convert them to sales locations. To add this data will be mandatory for a final version of the tool.

### *Regiotaxi*

The regiotaxi initiative once started as an alternative public transport modality. In (mostly rural) regions where the bus was not economically viable anymore, people were 'compensated' with the regiotaxi. This taxi service could be booked for a (fixed) price which was close to the price payed for a bus trip. The service has a few main differences with a regular taxi:

- The trip needs to be booked at least an hour before the time of departure;
- The pickup-time has a buffer of 15 minutes before or after the booked time;
- Trips can be combined, so you can end up with multiple people in one taxi;
- When a combination-trip is possible the taxi driver has the right to make a detour;
- The price of the trip is fixed.

Later on this initiative was also used as a means of subsidized transport for elderly and people with a handicap (WMO-vervoer). This extension of the service was so successful that 85% of the trips made with regiotaxi in the end were WMO related. In July 2017 the regiotaxi service in Twente was cancelled. Some municipalities did restart the service, but now only for WMO. The original purpose of the service, alternative public transport, doesn't exist anymore.

As data is managed by the province of Overijssel and this organization as PT authority is partnering this PDEng project, all trip reservation data from the period December 2013 till January 2016 has been made available for research purposes. This means from all the trips in this time frame among other data like the user-id, time of departure, and location of departure and location and time of arrival are available to construct (aggregated) OD-matrices.

### *Demographics*

Data from CBS, the Dutch centre for statistics, has been used to get data on the demographics of the postal code zones used in this research. As only data from 2010 is freely available on PC6 level, this data has been downloaded and used in this research as the data to base the demographical structure of the area on.

Data on this level of detail for more recent years is only available at a cost. For the level of PC4 the data is available openly. To have a more recent dataset on the PC6 level, regressions could have been used to update the 2010 PC6 dataset to the levels of 2017 to fit the changes which can be found on PC4 level. This exercise has not been done in this report as it wasn't the focus.

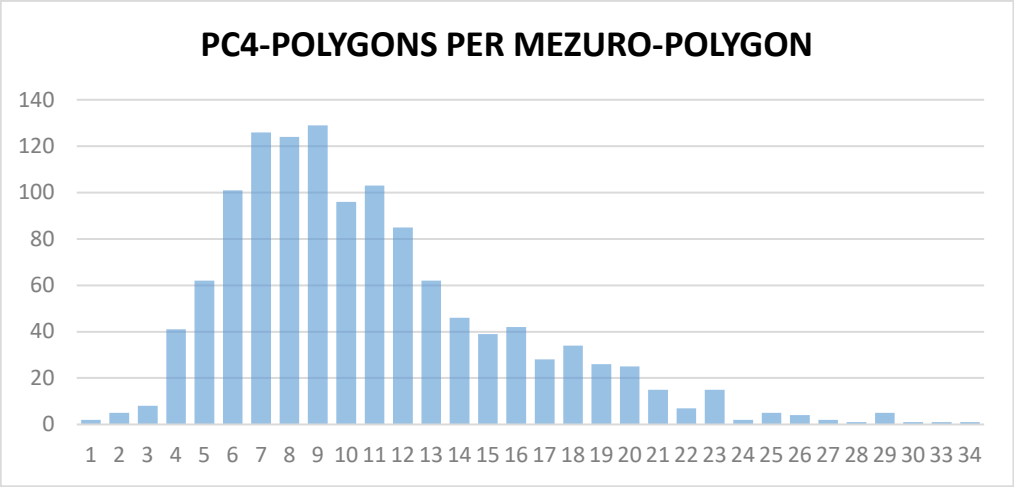


*Figure 4: PC4 - Mezuro comparison Netherlands as a whole*

### *Mezuro*

The company Mezuro uses mobile phone billing data to construct origin-destination tables. This data is provided for by the provider Vodafone. After anonymizing this data, it is sold to interested parties. This is potentially very interesting data, as it gives insight in real travel patterns for about 30,8% of the inhabitants in an area as this is the penetration grade of Vodafone in the Netherlands in 2018 (Kepinski, 2018). Using some smart algorithms it is even possible to trace back the modality used for every travel by using speed and route information. Research was done into the usefulness of this data source to determine if it is worth investing in for this project.

| Scale | Polygons |
|---|---|
| Mezuro | 1.243 |
| PC4 | 4.066 |
| PC6 | 449.839 |

*Table 3: Polygon comparison data sources*

The conclusion of this analysis is that for this project the data suitable for this project. As research in this project is done into creating OD-matrices on a low level scale (PC6 preferably) data needs to be available on a comparable scale. The scale for the Mezuro data is not even on PC4 scale (Table 3) and therefor, although very promising and interesting, deemed not usable for this project. Next to this scaling problem also the form factor of the polygons would be hard to match with the PC-format, as the mean Mezuro polygon overlaps with 10.8 PC4 polygons (Figure 4).

In the future the Mezuro data source may become more useful for research in smaller zones as at the moment the size of the zones depend on the scope of 4G masts which offer service over quite a large distance. The upgrade to 5G will offer masts with a smaller service area, which in turn will lead to ‘better’ OD-matrices. Future research in the field of travel data analysis, also on a small scale, should definitely include a new research into the evolution of this data source.

### *OViN*

OViN (Onderzoek Verplaatsingen in Nederland) is a multi-year still ongoing survey in the Netherlands which asks its respondents about travels made. This well respected survey instigated by the ministry of infrastructure and environment is one of the largest of its kind worldwide. With over 160.000 trip records it can be classified as a big data set. Although if the data is filtered to trips on the lower geographical levels (municipalities and / or modalities) only a few trips will be left to analyse. The OViN is available free of charge it is used as reference data to check if data sources used in the tool are in the same range.

| Min. | percentage |
|---|---|
| 0 | 33,07% |
| 1 | 1,37% |
| 2 | 2,74% |
| 3 | 2,55% |
| 4 | 1,42% |
| 5 | 33,07% |
| 6 | 1,39% |
| 7 | 2,01% |
| 8 | 1,62% |
| 9 | 0,81% |
| #NULL! | 19,95% |
| **total** | **100,00%** |

*Table 4: Reported travel time by last number of minutes*

Although this is a good source to use it has its limitations. As it is a survey people report behaviour instead of measure factual behaviour. This becomes very clear when the data is being analysed on one of the reported parameters; the travel time. In Table 4 only the last number of reported minutes was analysed. 18 became 8 and 33 became 3 for example. This was done to test if there is a bias in people to report on 'round' numbers like 0 and 5. According to the results there definitely is a significant bias towards reporting on the numbers ending with 0 or 5. If all data is taken into account, so errors as well (which account for +/- 20 percent of the data) 66% of reported minutes are on minutes ending with a 0 or either a 5. Without errors the bias towards 0 or 5 would even be 39,7% (33,07 + (0,3307*19,95)) per number, where 10% (2 out of 10 possibilities) would be the expected value.

### *Geographical vector polygons*

As this study is aimed at studying travel relations between different locations and making it easy to present the results, vector files with the boundaries of these regions are necessary to be able to plot the studied areas on a map. These polygons, or to be more precise the calculated centres of these polygons, can be used to calculate distances using the haversine-formula (Robusto, 1957). These distances can for example be used in a decay function which is to be used in estimating the attraction value of certain areas.

As most data in the Netherlands is collected on the level of postcodes, this study uses the postcode 4 (PC4) and postcode 6 (PC6) boundaries. Data not available on these levels will be distributed, in the transformation phase, to fit the postcode boundaries. It was also considered to use a 100m² grid file which CBS uses in a freely available demographical data file, but as most data is available in PC-format a lot of redistribution had to be done, which would have led to a loss of accuracy.

A PC4 polygon vector file by ESRI was obtained (Imergis, 2017), which is published on a regular basis under the CC-by license.

As a postcode 4 file was quite easy to obtain, the expectation was that this would also be the case with a PC6 file. While searching for a proper PC6 polygon / vector file it became clear this is data sold by private companies for a few thousand euros per data set, so it is basically not data which is publicly available under an open license. Although using a workaround the data has been downloaded from the internet from a database hosted by the university of Groningen. Also a file by Geodan was available in the archives from the university from a previous research. This last one is a licensed file and thus can be used for research purposes only.

### *KNMI*

The Dutch meteorological institute KNMI offers historical data on the weather on an hourly basis for 50 stations around the Netherlands. This source can be used to find a connection between patronage and for example rainfall by using a regression analysis. Although this is a very interesting feature to implement in the tool, because of time constrains this has not been implemented in the tool. The data has been put in the data warehouse, but no analysis is being done with it. Using this data in future development of the tool is strongly recommended.

### *Social media*

As we are creating a database filled with data from different sources, it is interesting to explore the possibilities of freely available social media data. If there is enough geo-located content available, it should be possible to create origin destination data out of social media posts as well which could be used as a source for the application.

After meeting with social media data expert dr. ir. Maurice van Keulen (EWI-UT) it became clear the amount of useful social media entries would be very low. As for trustworthy OD matrix generation you need regular users with at least 1 or 2 social media posts a day, with the geo-location option activated.

A recent study by Gkiotsalitis (in prep) for example confirms the assumption that social media platforms aren't an easy source for creating origin destination profiles as in this study for the whole metropolitan area of London only 32 profiles were suitable for analysis. Assuming the use of social media is comparable between Londoners and 'Tukkers' (people from the region of Twente), it means only about 2 or 3 profiles will be suitable for analysis (8,5 mln greater London area / 620.000 Twente area). This is way too little for a representative study and the reason why social media location data is not used in this study.

A more feasible possibility with social media data is using machine learning techniques to do text analysis. This could be used to measure sentiments on public transport in a certain area. This can be the objective of a different research project which complements this one as it is quite far from the intended goal.

### *Base map*

The tool needs to be easy in use that is why a graphical interface is being developed. As a base for this interface a digital map is used. Different providers offer free maps, the two biggest being Google maps and OSM (open street maps). Both can be used directly in projects but Mapbox makes it possible to easily integrate different maps, based on OSM into the web application frame work which is used in this project.

### *GTFS time table data*

GTFS (General Transit Feed Specification) is a file format developed by Google, in which time table data is being shared for use in travel applications. This data can be used to get insight in the frequency and route of public transport. This data is freely available on the internet. The data can be used to combine supply with demand. Out of time constraints the data is not used in the project.

### *Bus stops*

As OVCK data is essentially based on bus stops, it is also necessary to be able to show bus stops in the application. Bus stop locations have been derived from an internal table at Keolis as well as from an export of the open street maps data (Geofabrik, 2018).

### *Jobs*

For the amount of jobs available in a PC6 zone an export from the LISA (LISA, 2018) database has been used. This database consists of all companies in the Netherlands combined with the amount of jobs they offer. Depending on the purpose of the research this data can be bought on different geographical scales. The data is payed for per result. So an export on pc6 level will be more expensive than an export on pc4 level. Also an export with a minimum amount of 1 job per area is more expensive than a minimum of 10 jobs per area.

With this in mind an export of the region of Twente on PC6 level was bought by (then) Syntus in the Neolis programme to analyse the region of Twente with the intent to be complete but for an acceptable price. This is why the minimum amount of jobs was set on > 10. This gave a result of 3295 records. The amount of 10 was chosen as areas with less jobs were deemed uninteresting for analysis.

### *Conclusion*

To sum it up for the data selection chapter; from the thirteen sources considered, eight sources were deemed useful for this research;

- OVCK
- Regio Taxi
- Demographics

- OViN
- Geographical vector polygons
- Base map
- Bus stop data
- Jobs

This doesn't disqualify the usefulness of the other sources considered. With more effort or a different research question, the other sources can also have an added value in understanding travel behaviour.

### Data Pre-processing and transformation

In the previous section the use of data sources was described along with the considerations on whether to use them or not. In this section the steps used to transform the data which was deemed useful for use in in the previous section, are being described. The theory states this has to be done in two steps, data processing and later on data transformation. In practice it was easier to combine these two steps.

Formally the data pre-processing step in the data model involves six sub- steps (Han et al., 2011);

- cleaning
- Instance selection
- normalization
- transformation
- feature extraction
- selection

In this report no distinction is being made in these sub-steps. The pre-processing is described as one all-encompassing step and combined with the transformation step to prepare the data for use in this project.

#### OVCK

The OV-Chipkaart data is available at Keolis in a MSSQL-database. This database contains all OVCK-actions in all Keolis' concessions. For usability purposes this database already is cleaned and enriched by several ETL-processes to weed out as much errors as possible. This doesn't mean it isn't necessary anymore to clean the data as this is already done by the providing party, as there are still some errors left.

In essence this database shows the identifying number of the RFID-card, time of interaction, type of interaction, validator number and location of interaction (Table 2). By using algorithms this data is enriched to identify if it was a check-in, or check-out action, to identify the travel distance and some other parameters. As this database contains millions of records, an aggregating SQL query has been designed to limit the amount of data and processing needed in the application. As in the application it is not necessary to know every trip made, but only the OD-relation within a timeframe.

For this a bus stop matrix query has been designed which aggregates data from the OVCK database.

```
SELECT StationId, StationIdOut, hour, day, month,
year, Count(*) as aantal FROM (

     SELECT  [TransactionBusinessDate]
      ,[RouteId]
      ,[RouteIdOut]
      ,[StationId]
      ,[StationIdOut]
      ,[Concession_Code]
       ,[TripTimeDepartureIn]
       ,datepart(hh,[MsgReportDate]) as hour
       ,datepart(dd,[MsgReportDate]) as day
       ,datepart(mm,[MsgReportDate]) as month
       ,datepart(yy,[MsgReportDate]) as year

     FROM [Database_name] AS Transactions
     WHERE TransactionBusinessDate BETWEEN '2017-
01-01'  AND  '2017-12-31'  AND  TransactionType  IN
(30,32,34) AND Concession_Code = 'TWE'
```

```
) a

GROUP BY StationIdOut, StationId, hour, day, month,
year
```

This query provides a table in which for a chosen period, the amount of passengers traveling between two bus stops in a particular hour block in which the journey has started. With this table OD-travel matrices can be constructed.

*Trips and journeys*

The database contains trips. A trip can be defined as the time between a check in and check out action. A journey however also includes the connecting trip. For two sequential trips to be considered a journey, an interval of 35 minutes is common in the Netherlands, as this is also the time used in the OVCK business rules for having to pay a new boarding fee.

A very short trip  can disturb the OD-matrix as if the patron travels back to the origin within 35 minutes this will be counted as a 'round-journey' thus not including the destination where they were a short time. On a sample with 10.000 different patrons the difference in the amount of journeys made is 101.574 - 101.496 = 78 journeys or only 0,08%.

To calculate the journeys a script has been developed to check if a trip connects to a new trip within 35 minutes and to weed out the 0,08% of round-trips, a check was made to see if the destination of the connecting trip wasn't the same as the origin of the first trip. A version of the script can be checked below. In this version the journeys, journey time and amount of hops in the journey are being outputted. A later version writes the data to a journey table in the database out of which new OD-(journey)-matrices were created using the SQL mentioned in the previous chapter, about OVCK.

```
<?php
error_reporting(E_ALL);
$host = 'localhost';
$user = '***';
$pass = '***';
$database = 'pdengdb';


$db      =      mysqli_connect($host,      $user,
$pass,$database);

$query   =   'SELECT   MediaSerialNumberId   FROM
ovck_twe_raw GROUP BY MediaSerialNumberId';
$result = mysqli_query($db,$query);
echo mysqli_num_rows($result)."<br/>";$tot =0;
while($record = mysqli_fetch_assoc($result))
{
     print($record['MediaSerialNumberId'].'<br/>')
;

     $query1              =              'SELECT
StationId,StationIdOut,MsgReportDate  as  time_in,
MsgReportDate_Co  as  time_out,  TransactionType,
TransactionType_Co,  Bron  FROM  ovck_twe_raw  WHERE
TransactionType = 30 AND MediaSerialNumberId = '.
$record['MediaSerialNumberId'].' ORDER BY time_in
ASC ';
     $result1   =   mysqli_query($db,$query1)   or
die(mysqli_error());

     $records = mysqli_num_rows($result1);
     $j_stop_in = 0;
     $j_stop_out = 0;
     $j_time_in = 0;
     $j_time_out = 0;
     $hops = 0;
     $i = 0;
     echo 'Records: '.$records.'<br/>';

     while($record1                              =
mysqli_fetch_assoc($result1)){

          if(strtotime($record1['time_in'])     -
(35*6) < $j_time_out && $record1['StationIdOut'] !=
$j_stop_in){ // next checkin within 35 minutes AND
stationIdOut not the same as stationIdIn for the
first leg of the trip (quick roundtrips)
               $j_time_out                      =
strtotime($record1['time_out']);
```

```
                $j_stop_out                    =
$record1['StationIdOut'];
                $hops++;
                $i++;
            }
            else{
                if   ($i   !=   0   &&   $i   !=
$records){echo  $i.'  |  '.  $j_stop_in.  '-'  .
$j_stop_out . ' | Duration: ' . round(($j_time_out
-  $j_time_in)/60,0)  .  '  |  Hops:  '  .  $hops  .
'<br/>';}

$j_stop_in = $record1['StationId'];
$j_stop_out = $record1['StationIdOut'];;
$j_time_in = strtotime($record1['time_in']);
$j_time_out = strtotime($record1['time_out']);
$hops = 0;
$i++;
}
if($i == $records || $records == 1){

                echo $i.' | '.$j_stop_in. '-' .
$j_stop_out . ' | Duration: ' . round(($j_time_out
- $j_time_in)/60,0) . ' | Hops: ' . $hops . '<br/>';
            }
     }
}
```

This table then was used to create an OD-matrix which is journey based with an hourly basis;

```
SELECT station_id_in,station_id_out, HOUR(FROM_UNI
XTIME(time_in)) as hour, DAY(FROM_UNIXTIME(time_in
)) as day, MONTH(FROM_UNIXTIME(time_in)) as month,
 YEAR(FROM_UNIXTIME(time_in)) as year, count(*) as
 aantal
FROM ovck_journeys
GROUP BY station_id_out, station_id_in, hour, day,
 month, year
```

Regiotaxi

The regiotaxi data contains address-information which could potentially lead to privacy issues. To prevent this, all data has been scrubbed of any street and address information. Only the postal code was left as an identifier.

After removing corrupt and double files 25 csv files were imported in the data warehouse. This data was cleaned by removing cancelled trips. Bringing the total amount of usable records in the regiotaxi data to 1,096,876.

Steps that were taken to transform this data consisted of;

1. Identifying usable Excel sheets and removing duplications;
2. Transforming Excel sheets into .csv-files;
3. Adapting a script that handles importing .csv-files into MySQL-databases.

*Demographics*

Information about the amount of inhabitants, demographical composition, buildings, etc. was downloaded from the website of the Dutch bureau of statistics (CBS). For PC4 postcode polygons every year an export with this data can be downloaded for free. A recent export of the same data on PC6 level is harder to obtain. A free version of this file is only available from 2010.

For a production environment ideally the most recent export is bought so the data will match the most recent situation, as the travel data does as well. A cost effective, but time consuming, approach to update the 2010 data file is to do a regression analysis on the available pc4 data and copy these regression variables onto the PC6 file to get an updated file or redistribute the PC4 data on the PC6 polygons by using regression variables. Because of time constraints this last described exercise was not completed.

Next to CBS-data also data on amount of buildings and addresses was obtained from the polygon file by Geodan which will be described below.

### *OViN*

The OViN research file consists of 1 single excel file which consists of about 115.000 rows and 166 columns with data about travel preferences which have been obtained from people all over the Netherlands by using a continuous survey.

To import this file into the MySQL database the data was transformed into a .csv file. Using a variation of the php-script used to import the regiotaxi data, a database-table was created using the headers from the *.csv file which was then filled with data. By automating the creation of the database a lot of time was gained, as 166 columns didn't have to be created by hand, as was done before with the topography names which will be described later on.

### *Geographical vector polygons*

#### PC4

The postcode 4 file which encompasses the whole of the Netherlands in 4782 polygons, was downloaded as a shapefile (ESRI, 2018). To be useful in this project, some transformations had to be done to this file;

The shapefile consists of 40 MB of data, which gives a nice level of detail to the polygons, but the size is quite a large load when used in a browser environment. Also for the purpose of this project, the level of detail the file offers is not necessary, that is why the file was simplified. Next to being very detailed, the file is in the Amersfoort RD coordinate format, which is a Dutch coordinate system. This geographic coordinate system is quite uncommon in an international context and thus very hard to use in web development unless a transformation script is being written. Lastly the file format, .shp, is not very usable when developing a browser application. Ideally data is parsed as .(geo)json, a lightweight and flexible file format based on JavaScript.

As the API's are used to show the base map of the application, it is necessary to store data in a format which is compatible with the coordinate systems used by this provider. The shapefile had to be transformed to fit the coordinate system used in the base map. For this transformation ESRI ArcMap 10.0 was used. The coordinate system of the postcode 4 polygon file was altered to WGS 1984, a coordinate system which is the reference system for GPS (UNOOSA, 2018). This coordinate system is also in use as the most common format in cartographic web development.

To lower the file size of the application and get some gain in performance, the complex polygons had to be simplified. As the goal of this part of the project is to present data, and no calculations are being done using the postcode polygon dimensions, it is not a problem when polygons are not as detailed as they are in reality. To simplify a polygon, vertices (corners) need to be removed. There are different methods and algorithms to automate this process, the Visvalingam / weighted area algorithm was selected for simplification of the polygons as it promises a smoother appearance then the other available algorithms (Bostock, 2012). For a quick result the file was uploaded to mapshaper.org, where it is easy to simplify polygons as it shows the polygons on a map on which by using a slider it directly shows the new level of detail. The application this web site offers also makes it easy to choose different types of output, under which .json.

Once the .json file is prepared it has to be loaded into the database as single row polygons. To be able to import this file into a database a conversion script was designed. There is software available to do this, but these have licencing costs, whereas it is not too hard and quite educational to code a conversion script with the help of some openly available libraries. First the .json was transformed into a PHP array to be able to do manipulations for every single polygon in the data.

MySQL, as most modern database software, is able to store polygons native in a geometry object. This has advantages over storing them as plain text, such as being

able to use specific database functions to calculate distances between polygons, area sizes etc. Next to these benefits, it is good practice to store data in the proper format as optimisations are made in the database software for faster accessing data in the proper file formats and it is possible to create proper indexes on them which increases the speed of search queries and calculations dramatically if used well.

For the conversion from polygon information in the .json file to WKT (Well Known Text) which is used in a query to populate the database with geo-information, the open source library geoPHP was used. The resulting SQL-query was pushed to the database.

```
$file = file_get_contents('PC4NL.json');

$json_arr = json_decode($file,true);
foreach($json_arr['features'] as $key => $value){

     $query = '';
     $id = $key;
     $type = $value['geometry']['type'];
     $pc4 = $value['properties']['PC4'];
     $wkt_poly = json_to_wkt(json_encode($value));

     $query     =     "INSERT     INTO     geotest
(id,pc4,type,poly)                              VALUES
('".$id."','".$pc4."','".$type."',GeomFromText('$w
kt_poly'))";

}
```

The process as described above could probably be easier if other software was used. Research showed ArcMap 10.2 would be able to output .json, simplify polygons and probably can connect to MySQL directly. As this version of very expensive software was not available, the method as described was used, as it was free of charge.

### PC6 Free to use

The university of Groningen offers a 718 MB file with almost half a million polygons for free under a “CC-BY met vermelding van Esri Nederland, Kadaster” license (Groningen, 2016). To be able to download this file an account is needed for a closed and payed group moderated by Esri / ArcGIS the Netherlands. However, the service is in a crippled form also openly available from a query service which serves a maximum of 1000 records / polygon-vectors. The output is available in different file formats, including geojson.

As it would take 456 manual actions to retrieve all data from this service, a script was developed to automate this process and in the same run populate the database with the polygons with the help of the geoPHP library;

```
<?php

include_once('geoPHP/geoPHP.inc');

function json_to_wkt($json) {

  $geom = geoPHP::load($json,'json');
  return $geom->out('wkt');
}

$mysqli  =  new  mysqli("localhost",  "***",  "***",
"pdengdb");

if ($mysqli->connect_errno) {
    echo "Error: Unable to connect to MySQL."

    exit;
}
$loop = 500;
$time_start = $_SERVER['REQUEST_TIME'];
for ($n = 0; $n < $loop; $n++ ){
      $offset = $n+1000;
      $file                                    =
file_get_contents("https://geo.rug.nl/arcgis/rest/se
rvices/Administratief/PostcodegebiedenNL/FeatureServ
er/1/query?where=POSTCODE+LIKE+%27$offset%25%27&obje
ctIds=&time=&geometry=&geometryType=esriGeometryPoly
gon&inSR=&spatialRel=esriSpatialRelRelation&distance
=&units=esriSRUnit_Meter&relationParam=&outFields=&r
eturnGeometry=true&maxAllowableOffset=&geometryPreci
sion=&outSR=&gdbVersion=&historicMoment=&returnDisti
nctValues=false&returnIdsOnly=false&returnCountOnly=
false&returnExtentOnly=false&orderByFields=&groupByF
```

```
ieldsForStatistics=&outStatistics=&returnZ=false&retu
rnM=false&multipatchOption=&returnTrueCurves=false&
sqlFormat=standard&f=geojson");

      $json_arr = '';
      $json_arr = json_decode($file,true);


      foreach($json_arr['features']  as  $key  =>
$value){
            $type = '';
            $pc = '';
            $wkt_poly ='';
            $query = '';
            $type = $value['geometry']['type'];
            $pc = $value['properties']['POSTCODE'];
            $wkt_poly                          =
json_to_wkt(json_encode($value));
            $query  =  "INSERT  INTO  poly_pc6
(pc6,type,poly)                          VALUES
('".$pc."','".$type."',GeomFromText('$wkt_poly'))
      ON DUPLICATE KEY UPDATE dup = dup + 1";
            if (!$mysqli->query($query))
                  {echo  "<br/>ERROR:  ".$mysqli-
>error;A
      }
}
$mysqli->close();
?>
```

This script revisits the results page for 500 times, every time offsetting the query pushed to this page by a 1000 results. Every iteration the polygons get pushed directly to the data warehouse.

### Limited data set

At the university of Twente a Geodan PC6 dataset is available which is usable for research only. This set has also been incorporated in the data warehouse as next to the polygons it contains interesting demographical information on amount of inhabitants and buildings which otherwise had to be derived from the PC4 datasets which would have led to inaccuracies in the data. The data can only be used in production when a license is bought, the dataset will therefor only be used for research purposes and comparison. This data will be removed when the tool is handed over to Keolis. This will have little impact as a replacement data file is available with the RUG dataset described before.

### Calculate area size and polygon centroid

As distances between polygons need to be calculated, a point within the polygon had to be defined as the middle of every single polygon. In regression analysis it can also be useful to know the size if the polygon, therefor also the area size was calculated. MySQL has native functions to calculate area sizes and polygon centroids when data is stored in a geometry format;

```
UPDATE poly_pc4 SET centroid = Centroid(geometry),
area = area(geometry)
```

Using this query the records in the database were updated with the centre as well as the area size. Although it is not following the rules of database normalization, updating the table with this static data is, as it only costs a few bytes in storage, far quicker than having to recalculate these values every time when the data is needed in a production query.

### Postcode topography names

A problem with data on the postal code level is that nobody can really relate to it. As most people do know and relate to place names, a table needed to be added to the database which could be used as a bridge between postal code and place names.

A script was created to push this ‘postcodetabel’ .csv file, which was available on some obscure website, into the database. As it was a rather small data source the headers were added by hand, as it took too much time coding it in the conversion script. Later on the conversion script was altered as larger sources were added to the database using the basis of the same script.

This data can be used as enriched information to the postcode table. It is for example possible to retrieve only the polygons in a certain province or municipality by joining this data on the postcode table by using the query below;

```
SELECT    poly_pc6.id,    poly_pc6.pc6    as    pc4,
poly_pc6.type,  ST_ASBINARY(poly_pc6.poly)  as  wkb,
pclocation.gemeente          as          gemeente,
pclocation.provincie as provincie
           FROM poly_pc6
           LEFT JOIN pclocation ON poly_pc6.PC4 =
pclocation.pc4
           WHERE gemeente = "Enschede"
```

*Base map*

As the base map is only used for presentation purposes it was not necessary to pre-process this data as it is used as is by making use of the API offered by Mapbox (Mapbox, 2018) which can be integrated in the application framework quite easily. More information about this can be found in the section about the development of the front end of the application later on in this report.

*Bus stop data*

The website Geofabrik (2018) offers data exports from open street maps (OSM). These files contain information in shapefiles about all geographic features available in OSM. These files can be used in every common GIS (geographic information system) to manipulate and analyse the features contained in the files.

Using arcgis the bus stops were extracted from the file, saved as .csv with gps coordinates and imported into the database where the GPS-coordinates were transformed into a geometry object.

Next to this openly available data also an export from the Keolis database was used in which bus stops with identification number and GPS-location was available.

*Jobs*

The jobs from the LISA database (LISA, 2018) was available as a .csv file with only two columns, a PC6 and the amount of jobs in this postalcode zone. This csv could easily be transferred into the database by making use of a slightly altered csv to database script which has also been used to load the regiotaxi data into the database.

### Data Transformation

As described in the previous section the step of data pre-processing was combined with the practice of data transformation and also the steps taken were explained per data source in this section. For the transformation part, two scripts were created which evolved a bit over time and of course were altered based on the data source.

*Json and csv-files to database*

The script below was used to push csv files to the database. It connects to the database, creates a table if it does not exist already and loops through the data to push it into the newly created table, creating not yet existing columns while looping through the data. A shortcoming in this script is that it does not check the type of data it pushes. All data is pushed in varchar format. Later on datatypes in the database have been changed manually to accommodate the underlying data better, improving querying times. Making this script more advanced this could of course be automated.

```
<?php
$host = 'localhost';
$user = 'root';
$pass = '';
$database = 'pdengdb';
$delimiter = ';';


$db = mysqli_connect($host,$user, $pass,$database);
```

```
if (!$db) {
    echo "Error: Unable to connect to MySQL." .
PHP_EOL;
    echo      "Debugging      errno:      "      .
mysqli_connect_errno() . PHP_EOL;
    echo      "Debugging      error:      "      .
mysqli_connect_error() . PHP_EOL;
    exit;
}

$file = 'filename.csv';
$table = 'databasename';


$fp = fopen($file, 'r');
$frow = fgetcsv($fp,0,$delimiter);
$columns = '';
foreach($frow as $column) {
    if($columns) $columns .= ', ';
    $columns .= "`$column` varchar(20)";
}

$create = "create table if not exists $table
($columns);";

echo '<pre>'.$create."</pre>";
if(!mysqli_query($db,$create))
     {echo("Error      description:      "      .
mysqli_error($db));}
```

This script was used as well for importing other data like the OVCK matrix data or the OViN files into the database which was available or transformable into .csv.

To import (geo)json files into the database a different script was developed:

```
<?php
include_once('geoPHP/geoPHP.inc');

function json_to_wkt($json)
  $geom = geoPHP::load($json,'json');
  return $geom->out('wkt');
}
$mysqli   =   new   mysqli("localhost",   "admin",
"passwd", "pdengdb");
if ($mysqli->connect_errno) {
    echo "Error: Unable to connect to MySQL.";
    exit;
}
$file = file_get_contents('GEOJAY.geojson'
$json_arr = json_decode($file,true);
foreach($json_arr['features'] as $key => $value){

     $query = '';
     $id = $key;
     $type = $value['geometry']['type'];
     $pc6                =                $mysqli-
>real_escape_string($value['properties']['PC6']);
     $pc4                =                $mysqli-
>real_escape_string($value['properties']['PC4NR'])
;
     $addr               =                $mysqli-
>real_escape_string($value['properties']['AANT_ADR
ES']);
     $pand               =                $mysqli-
>real_escape_string($value['properties']['AANT_PAN
D']);
     $wkt_poly = json_to_wkt(json_encode($value));
     $query     =     "INSERT     INTO     new_pc6
(geometry,pc6,type,pc4,panden,adressen)     VALUES
(GeomFromText('$wkt_poly'),'".$pc6."','".$type."',
'".$pc4."','".$pand."','".$addr."')";
     if (!$mysqli->query($query))
           {echo     "<pre>ERROR:     ".$mysqli-
>error."</pre>";
           echo "<pre>".$query."</pre>";}

     flush();
     ob_flush();
}
$mysqli->close();
?>
```

This script loops through the json file and puts the values in a manually prepared table. GPS-locations are being transformed into native SQL geometry objects using a freely available php-library; GeoPHP.

### Data Mining / results

This section shows the result of different analysis which were done on the prepared data to gain more insight in travel patterns of people using the systems for which

data is available. Also analysis were done which don't relate directly to the main goal of this research (the creation of a DSS). These analyses are also shown here to give insight in the possibilities of constructing information on travel patterns with the data available in the data warehouse. Later on this information could be in some form integrated into a further developed version of the DSS.

### *Decision Support System*

Most energy has been put into the creation of a decision support system in which data from the database can be shown on postcode level. The user of the system can, using this system, do their own data mining by selecting a polygon / selection of polygons and be able to see the interaction from this polygon with other polygons. More on the development of this system can be found in the section about web development later on in this report.

### *Distance analysis OVCK service points*

Based on street data, which was derived from openstreetmaps, an analysis on distance between OVCK service points was conducted. At these service points, which are mostly supermarkets and convenience stores, the OVCK can be loaded with travel credit or subscriptions. Keolis had a question on the geographical spread of these service points.

Using internal data on the exact locations, combined with street data from OSM using ArcGIS, analysis were done to determine the walking distance between these service points by creating an origin destination matrix between all points.

### *OVCK data analysis*

For more insight in travel behaviour of patrons in Twente some analyses have been conducted which will be explained in this chapter.

#### Based on trips

In Figure 5 the amount of trips in the whole of the region of Twente made per hour block (trips starting in that particular hour) are shown for the month March for two different years. What is noticeable in this chart is that there is almost no change in travel pattern visible between the two years. The morning peak is a little bit higher in 2018 than it was in 2017 and the peak of the evening rush seems to be an hour(-block) earlier. Numerical differences are small though.

In Table 5 it is noticeable that there is no real difference in the most important bus stops between 2017 and 2018. The only difference is the disappearance of Haaksbergen Busstation in 2018. This can easily be explained as this bus stop

wasn't serviced during the reconstruction of the Eibergsestraat which was done in this month. A symmetry analysis on the use of the 25 most used bus stops can be found in Table 6. The difference between the amount of boarding and alighting passengers is rather small. Only a few stops stand out with a difference > 20%; 40406 (30%, Enschede, Kennispark) and 40033 (23%, Almelo Centrumplein). Both stops have more people leaving the bus at that stop than people boarding there. For both locations this is not really remarkable as Kennispark is the gateway from the regional train to the university. In some cases it could be that walking to the university is faster or as fast as waiting at the station for a bus. The other way around the waiting time can be done at the university, which delays the start of the trip.

For Almelo Centrumplein the effect can be explained by the motive of people going there; shopping. Start your shopping by going to the centre by bus, walk around in the centre and take a bus or train at the other side of the centre.

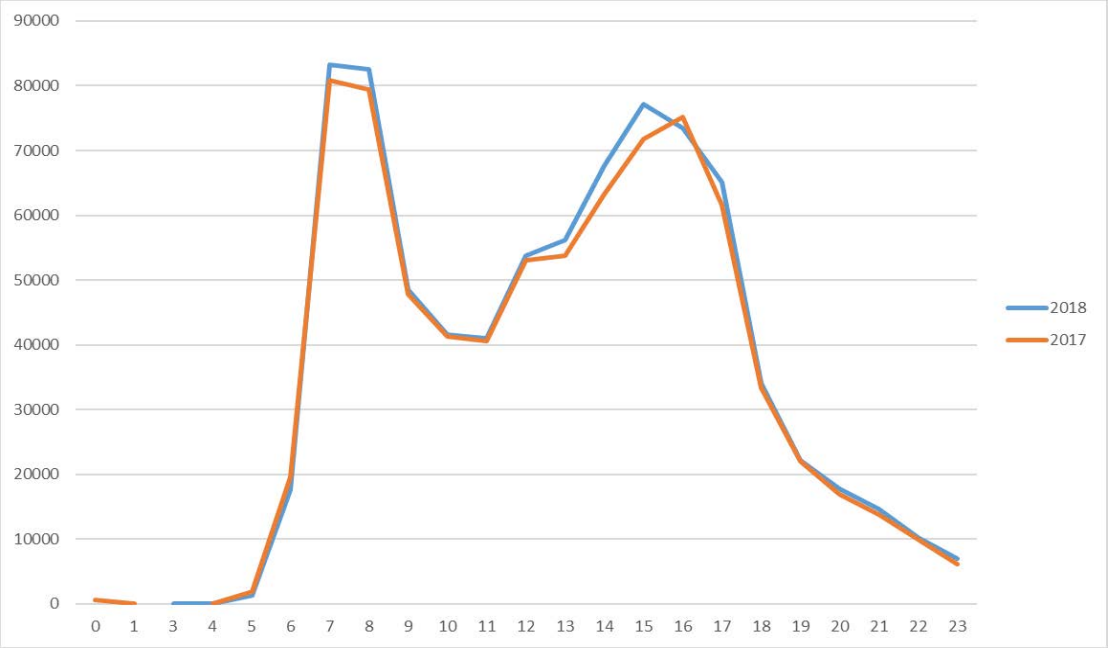


*Figure 5: OVCK trips per hour block March 2017 vs March 2018 Twente*

| | 2017 | | | 2018 | | |
|---|---|---|---|---|---|---|
| **rank** | **id** | **name** | **amount** | **id** | **name** | **amount** |
| **1** | 40300 | Enschede, Centraal | 155580 | 40300 | Enschede Centraal | 161576 |
| **2** | 40591 | Hengelo, Centraal | 68014 | 40591 | Hengelo Centraal | 74401 |
| **3** | 40032 | Almelo, Centraal | 48466 | 40032 | Almelo Centraal | 49377 |
| **4** | 40433 | Enschede, van Heekplein | 27862 | 40433 | Enschede, van Heekplein | 31919 |
| **5** | 40835 | Oldenzaal, Station | 23069 | 40835 | Oldenzaal, Station | 23092 |
| **6** | 40524 | Haaksbergen, Busstation | 12488 | 40518 | Goor, Station | 9803 |
| **7** | 40518 | Goor, Station | 8307 | 40445 | Enschede, Westerb… | 8507 |
| **8** | 40445 | Enschede, Wester… | 8271 | 40429 | Enschede, Utrechtlaan | 7312 |
| **9** | 40429 | Enschede, Utrechtlaan | 7342 | 40166 | Borculo, Busstation | 6443 |
| **10** | 40360 | Enschede, Kennispark | 6857 | 40360 | Enschede, Kennispark | 6179 |

*Table 5: 10 most used stations for departure 2017-2018 Twente*

| stopid | totalin | totalout | avg | diff | diff perc |
|---|---|---|---|---|---|
| 40300 | 161576 | 157729 | 159653 | 3847 | 2% |
| 40591 | 74401 | 64100 | 69251 | 10301 | 14% |
| 40032 | 49377 | 44975 | 47176 | 4402 | 9% |
| 40433 | 31919 | 31230 | 31575 | 689 | 2% |
| 40835 | 23092 | 22759 | 22926 | 333 | 1% |
| 40518 | 9803 | 9205 | 9504 | 598 | 6% |
| 40445 | 8507 | 8666 | 8587 | -159 | -2% |
| 40429 | 7312 | 7544 | 7428 | -232 | -3% |
| 40166 | 6443 | 6421 | 6432 | 22 | 0% |
| 40360 | 6179 | 6404 | 6292 | -225 | -4% |
| 40406 | 5189 | 6726 | 5958 | -1537 | -30% |
| 40541 | 5864 | 5556 | 5710 | 308 | 5% |
| 40394 | 5290 | 5375 | 5333 | -85 | -2% |
| 40762 | 5259 | 4410 | 4835 | 849 | 16% |
| 40740 | 4281 | 4271 | 4276 | 10 | 0% |
| 40905 | 3872 | 4402 | 4137 | -530 | -14% |
| 40304 | 3970 | 3884 | 3927 | 86 | 2% |
| 40033 | 3283 | 4030 | 3657 | -747 | -23% |
| 40447 | 3510 | 3606 | 3558 | -96 | -3% |
| 40453 | 3531 | 3408 | 3470 | 123 | 3% |
| 40379 | 3197 | 3686 | 3442 | -489 | -15% |
| 40355 | 3506 | 3301 | 3404 | 205 | 6% |
| 40459 | 3177 | 3203 | 3190 | -26 | -1% |
| 40301 | 3242 | 3088 | 3165 | 154 | 5% |
| 40422 | 2816 | 2994 | 2905 | -178 | -6% |

*Table 6: Symmetry analysis top 25 bus stops Twente*

| PC4 depart | PC6 destination | sum | location | Location type |
|---|---|---|---|---|
| All | 7534AM | 4.768 | Liberein Bruggerbosch | Nursing home |
| All | 7575EC | 2.334 | Top-Craft | Sheltered employment |
| All | 7651DH | 2.258 | Kringloopbedrijf de Beurs | Sheltered employment |
| All | 7554PG | 2.015 | De Tukkerij | Care farm |
| All | 7448PG | 2.005 | Zorgboerderij de Schurinkshoeve | Care farm |
| All | 7481EV | 1.929 | Livio | Care appartment |
| All | 7575EE | 1.529 | Kringloopbedrijf de Beurs | Sheltered employment |
| All | 7513ER | 1.336 | MST Enschede | Hospital |
| All | 7609PP | 1.320 | ZGT Almelo | Hospital |
| All | 7573AV | 1.300 | Station Oldenzaal | Station / sheltered employment |
| All | 8101ZW | 1.245 | De Zonnehof | Special education |
| All | 7491NP | 1.191 | Restaurant In de Hagen Delden | Restaurant |
| All | 7576AV | 1.143 | de Zonnestraal | Nursing home |
| All | 7447AV | 1.132 | ZorgSaam Hellendoorn | Health care center |
| All | 7481AV | 1.084 | J.P. vd Bentstichting | Elderly home |

*Table 7: OV-trips PC6 top 15*

Some analyses on travel patterns have also been done on the regiotaxi-data. A list of the most popular destinations have been generated on different levels by creating the ETL process in the figure below (Figure 6).

Based on the results of these tables the 15 most important PC6 locations visited by people not having a WMO (care) indication were looked up (Table 7). These locations could be interesting to provide a regular service to, as people already pay the (semi-)regular taxi tariff to get there.

Using the data available it is also possible to create user profiles;

*20.855 users in database*

- Top 25 users (0,12%) used 3,33% of the zones driven
- Top 5% users used 39,55% of the zones driven
- Top 20% users used 74,49% of the zones (testing 20/80 rule)

These results are interesting for providers of new types of regular public transport as they mean that about 1000 people are responsible for almost 40% of the travelled kilometres. A service focussed on a relative small part of the population could be profitable quite quickly this way. A recommendation is to delve deeper into this top 5% to check if in this group commonalities in travel patterns can be found. Are there clusters of users and times to be found in this user group which can be serviced in a regular service.

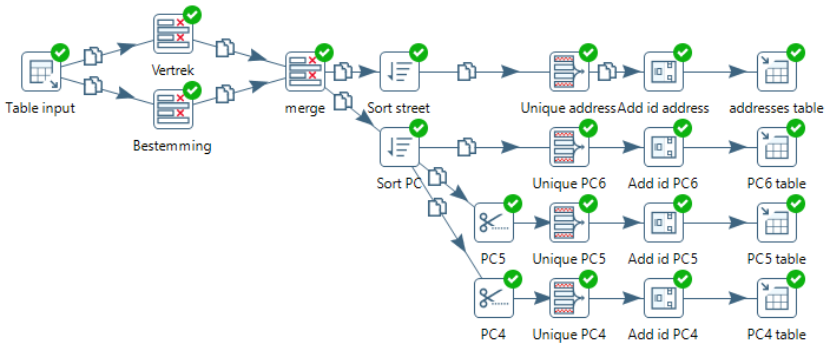


*Figure 7: Regiotaxi ETL process*

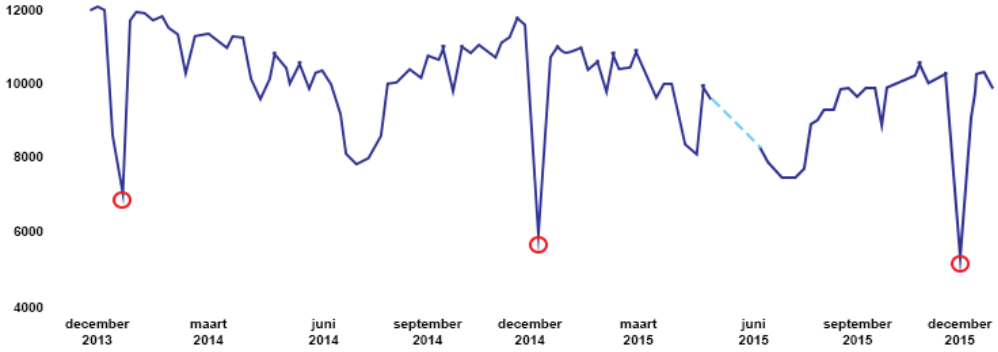


*Figure 6: Amount of regiotaxi trips made per day in Twente*

Travel patterns of people using the regiotaxi service differ over time, the amount of trips made for example can be seen in Figure 7. The dips in trips made which are circled in red are the Christmas days, the dashed line is a month of missing data. What else can be spotted is the trend that the use of taxi is biggest during the winter months and significantly lower during summer / vacation time (the dip in June '15 is due to missing data).

Another analysis which can be done is a look into the regular travel patterns of regiotaxi users. This is done by plotting the travel date against the average amount of days the trip was booked in advance (Figure 8). What is noticeable, except for the missing data (dotted line) is that the high peaks are the Christmas days and the low spikes every week between 30 and 40 days booking in advance, are the weekends.

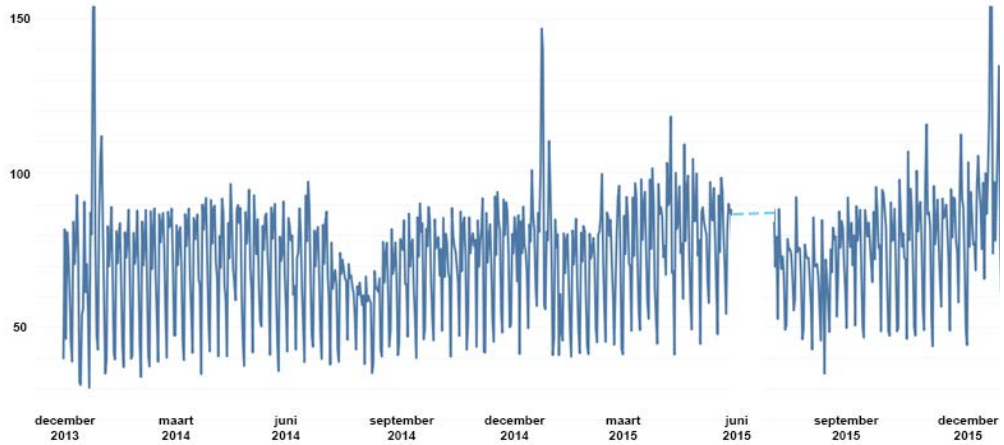


*Figure 8: Days before booking regiotaxi trip in Twente*

People using the regiotaxi thus have quite regular trips during working days (70/90 days in advance) and are more spontaneously booking a trip in the weekends.

More analyses on the same data source can be found in the report by Janssen (2017). In his report he for example concludes that regiotaxi users in cities mostly travel within the city and users from smaller villages tend to travel larger distances towards nearby cities. Another conclusion is that the OD matrix is more or less symmetrical, which means in an origin zone there are more or less the same amount of departures as there are arrivals. The report concludes with the notion that most trips made by regiotaxi are made to hospitals, city centres and sheltered work places.

### *Distance decay function*

With the data now available in the data-warehouse, distance decay functions as described in the theory section can now be calculated.

The observed data which is labelled as '*_COUNT' in Figure 9, Figure 10 and Figure 11, clearly display an S-curve. This is why the following formula is used to calculate regression variables α and β by minutes γ.

$$\frac{1}{1+\exp(\alpha+\beta\log(\gamma))}$$

In the analysis trips longer than 90 minutes were deleted from the data, as these influence the results and mostly will be bad registrations as there are almost no possibilities to travel this amount of time in a bus.

Using this method the following variables were calculated with the complete OViN database and the January and February OVCK data (Table 8):

Using these parameters different curves can be generated:

Some conclusions which can be drawn from this analysis are;

| Source | Selection | α | β |
|---|---|---|---|
| OViN | PT Nederland (inc. pre and post transport) | -11,567 | 2,907 |
| OViN | Car Nederland | -4,966 | 1,833 |
| OViN | Car Twente | -5,878 | 2,396 |
| OVCK | Keolis | -5,068 | 2,143 |
| OVCK | Twente | -6,696 | 2,667 |
| OVCK | Midden Overijssel | -5,251 | 2,324 |
| OVCK | Veluwe | -6,345 | 2,382 |
| OVCK | Provincie Utrecht | -5,749 | 2,273 |
| OVCK | Almere | -3,698 | 1,921 |
| OVCK | Morning rush (7-9) | -5,430 | 2,210 |
| OVCK | Evening rush (16-18) | -5,087 | 2,143 |
| OVCK | Total rush | -5,267 | 2,177 |
| OVCK | Off peak | -4,915 | 2,135 |
| OVCK | City bus (9; Enschede - Hengelo) | -7,289 | 3,061 |
| OVCK | Regional bus (62; Denekamp - E'de - Borculo) | -8,071 | 2,708 |

*Table 8: Distance decay parameters*

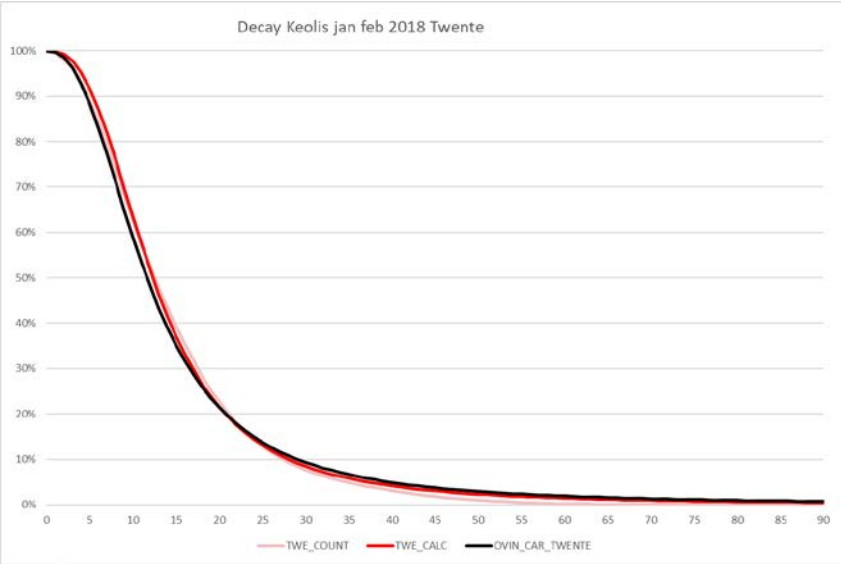


*Figure 9: Decay Twente OVCK bus / OViN car*

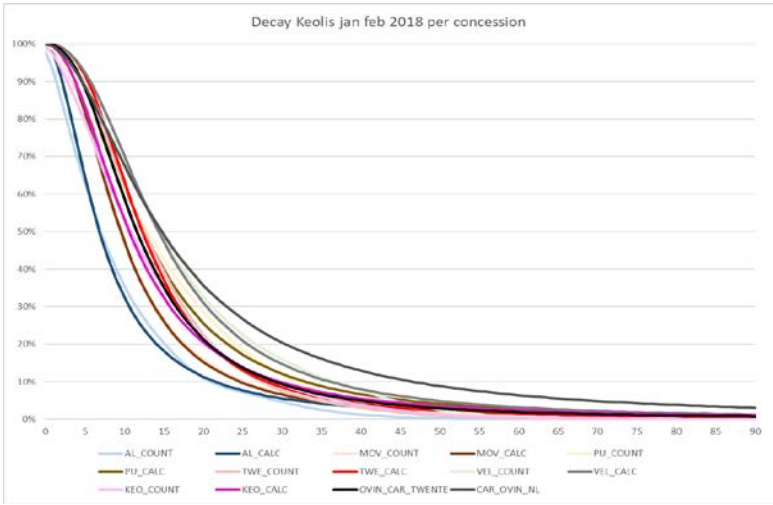


*Figure 10: Decay per Keolis concession*

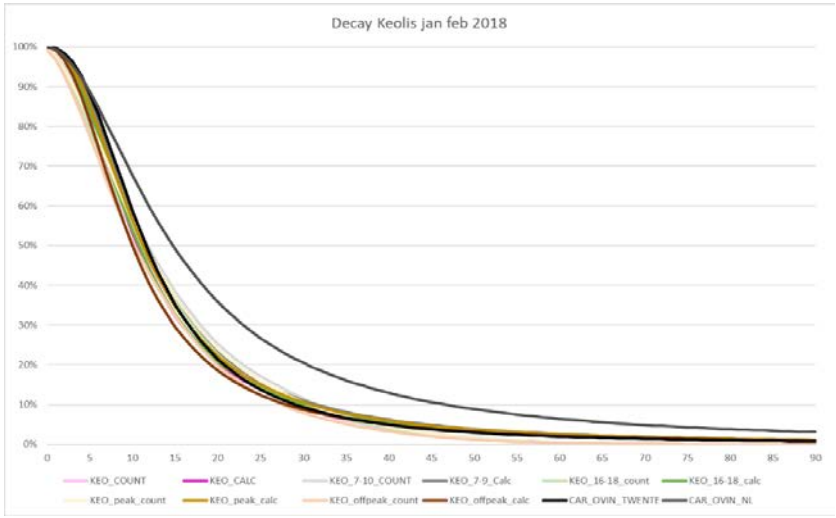


*Figure 11: Decay peak / off peak*

- In rural areas people tend to travel longer. In Almere (urban area) only 20% of trips made take more than 15 minutes compared to the rural Veluwe area where almost 50% of the trips made last more than 15 minutes.
- There are significant differences between the mean travel time by car in the Netherlands when compared to Twente.
- The Twente car curve is very close to the Twente OVCK curve.
- There is not a real difference between the time travelled between rush and non-rush hours.

### Use of distance decay analysis in practice

Combining these distance decay functions with data on modality travel times between zones and a quantifiable factor of the destination zone, the accessibility of these quantifiable factors can be obtained. Questions which can be answered are for example how many jobs (quantifiable factor) are reachable in a destination zone at different times of the day and/or with different modalities. Calculating this for

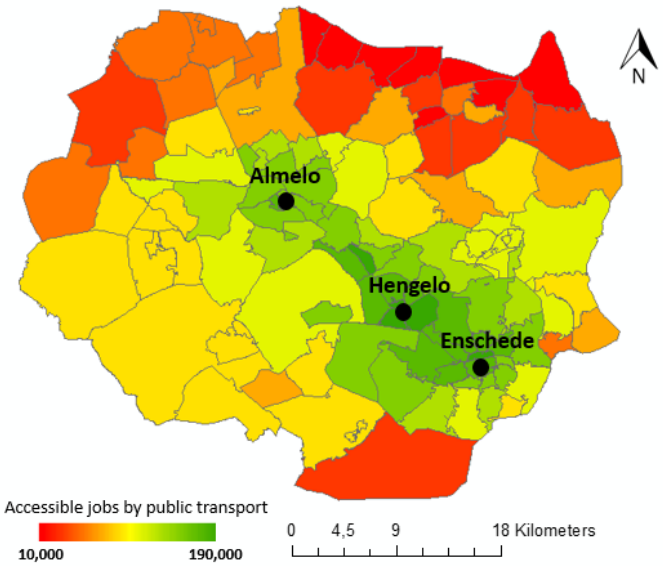


*Figure 12: Job accessibility by PT using distance decay*

different origins and a large set of destinations, gives the possibility to compare the origins on accessibility of services and rank them on attractiveness.

In Figure 12 an example can be found in which the decay function has been used in combination with job data and a mean of 4 consecutive 5 minute time slots of GTFS data outside of the rush hour, to calculate job accessibility within the borders of Twente on PC4 level. Not totally surprising the best job accessibility is in the three cities of Almelo, Hengelo and Enschede, where most jobs are and PT supply is best.

What also can be seen quite clear in this figure is the edge effect (Stewart Fotheringham & Rogerson, 1993), in which the potential of areas near the border of the researched area have a lower potential due to not taking into account the possibility for people to reach areas outside of the research area. In practice for example the amount of jobs reached for the polygons in the west of Twente will be higher due to the possibility to also reach a significant amount of jobs in Zwolle and Deventer within the distance decay curve. This effect will be much lower in the other directions, as in the east there are very few connections to Germany, in the north as well connections are lacking. In the south some connections exist, but they are to areas with a low amount of jobs (Achterhoek). A correction therefor should be mainly done for the areas in the west of Twente.

The results from a distance decay analysis can also be used in price differentiation for commercials shown in the buses for example. Most bus companies show commercials on screens inside. Using a distance decay function, the amount of time a passenger is exposed to these messages can be quantified. In combination with te available data on occupation the pricing of this commercial space can be differentiated on.

## (Web)development

### Design specifications

This chapter discusses the considerations and specifications which the artefact will be built on. The framework used was derived from Appleton (1997), who put together a flexible but complete software design specifications framework based on different methods used.

#### *Dependencies*

#### Assumptions

An assumption made is that it is possible to mine data from different sources into usable and reliable information. It could be the data which is available can't be presented in a way it provides usable information with more value than the surveys which are used now.

#### Related software

The artefact will make use of PHP and SQL and therefor interacts with a webserver and a browser. As this is standardized and openly available software it won't be a problem during the design process.

#### Operating systems

The main focus will be on desktop computer use. Windows within a Citrix environment. During the process the design for mobile devices will be taken into account while making interface design choices.

#### End-user characteristics

The end-users of the tool will be people involved in making plans for changes in level of service. Not data scientists. Later on, if the tool is also usable for other departments, more proficient users could be added to the list of end-users. Therefor it is important to keep the tool as simple as possible.

#### Possible and/or potable changes in functionality

The tool can be expanded by adding more data to give more insight into travel behaviour or patterns and locational factors as demographics.

Also the principles of this research can be used within other software packages which are or maybe will be in use later on in the company. An example could be the use of data which is part of this research within the report functionality of Microsoft Business Objects, which is the software package used for reporting within the company at the moment.

### *General Constraints*

#### Hardware environment

The interface of the tool should be lightweight, as it will have to run within a Citrix environment which is a server based windows environment. On the backend in theory more power can be used for data mining purposes, as this can be scheduled for periods when not much users are at work or extra computing power can be created or bought. Although the aim is to code as lightweight as possible.

#### End-user environment

There is a possibility the end users will stop using Citrix / windows. That is why the artefact will be programmed in PHP outputting html, which works independent of the operating system as it will run in any modern browser.

#### Availability of resources

Regiotaxi data has dried up, as the service stopped functioning in July 2017. Other sources which are used for the creation of information at the moment will be available for a longer period. Nonetheless it is important to consider the need for new and more sources of location data in the future.

#### Security requirements

Because of the use of real location data, it is very important to respect the privacy laws and ethics. No trip must be able to lead back to a single person.

Next to privacy laws and possible issues, the internal check-in data is competitive data which have to be stored safely, as competing businesses could take advantage of this during tenders.

### *Goals and Guidelines*

- Work on a regular desktop computer
  Because of the Citrix environment used within the company

- Be intuitive for the user
  As the intended users (in general) don't have much feeling with data science to adjust lots of parameters to refine the results.

- Support in decision making, not make decisions itself
  The end-users should be able to make their own decisions based on multiple sources and own intuition, knowledge and experience. The system shouldn't be intrusive in this process.
- Possible to integrate in the current data warehouse
  The data used for this artefact could also be very useful in other departments of the company. Therefor it would be nice if the data would be openly available for internal use, to be able to use it within other applications already in use within the company.
- Be more effective than current methods of research used at Keolis
  Methods currently in use for estimating the potential for new services are mostly intuition, ad hoc small scale data analysis and surveys. More effective in this sense means time, money and or reliability.
- Be expanded into a tool which encompasses more than just the current research and Twente region
  During development the tool will be focussed on a limited geographical area. For the future it must be possible to refine the tool and add more data to make it a useful source of information for more areas for example during tenders.

*Policies and Tactics*

Trade-offs

- The amount of data sources used
  Ideally as much data sources as possible will be used in the tool. Because of time constraints and costs associated with the acquisition of data, compromises have to be made on the amount of data sources.
- The amount of regions / types of communities which can be analysed
  Because of time constraints with the PDEng program the tool will not be developed as a general tool for all regions, but be limited for use in the Twente region.


*Figure 13: Conceptual design*

## Conceptual design

In Figure 13 a schematic of the conceptual design of the decision support tool is shown. Data is transformed using ETL-processes into usable tables. The user interface interacts with the backend by user defined parameters which will load the data from the database to be presented in the interface.

## Set-up

### Backend

*Server software*

For the development of the decision support system a WAMP package was used. WAMP is an acronym for Windows, Apache, MySQL and PHP. The windows part of this acronym refers to that the package is suitable for this operating system, LAMP offers the same functionality but for the Linux environment.

Apache is web server software which allows the use of PHP as a dynamic programming language to output proper html, which is the static page the end user is seeing in their web browser. To store the data used in this project, the open source database software MySQL was used. To set up, maintain and view the content of the database, phpmyadmin, a database management tool is used. For database diagnostics also the MySQL console windows software is used.

For the development of scripts, querying the database and outputting data, PHP is used. Coding is done in Dreamweaver. PHP is a programming language used in web development for the creation of dynamic pages. Another language used in this

project is JavaScript, which is a client side web language, by using i.e. the jQuery library it can be combined with PHP to also use data from the MySQL database.

## Frontend

### *Libraries*

#### geoPHP

The library which is used for transforming .json into WKT (Well Known Text) for insertion into the database is also used the other way around. SQL-queries which are constructed provide polygons in WKT which have to be transformed into .geojson. The geoPHP library does this by providing a simple function to transform one data type into another.

#### Leaflet

To present data on a map easily using points, lines and polygons the JavaScript library Leaflet is used. Leaflet offers functions which make it quite a lot easier to access different map provider API's which can be used as a background layer for the application. For the data which is collected for this project, Leaflet has functions which make it possible to present them nicely on top of the background layer by transforming the .geojson extracted from the database into objects which can be interacted with using JavaScript.

# Product development

## Database overview

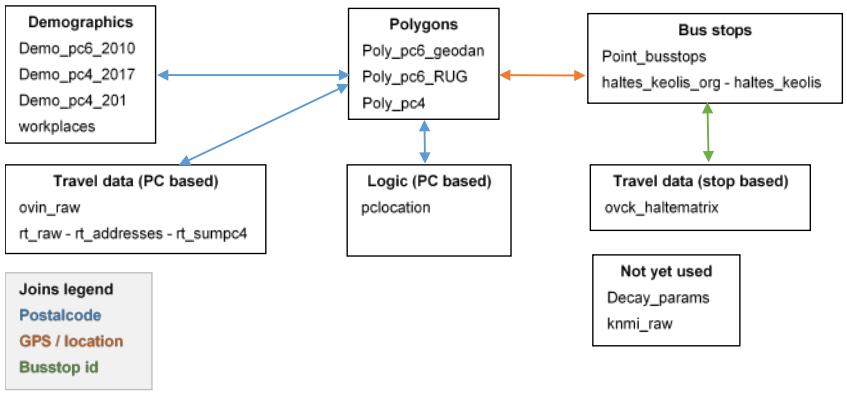


*Figure 14: Database overview*

## Programming

As described before, a mix of programming languages were used to build this application.

- The data is being queried by using SQL
- The design is being constructed by CSS
- The design is shown in HTML
- The interaction is handled by JavaScript
- The language to build the system on and glue it all together is PHP

Programming is done in the development environment of Adobe Dreamweaver.

## Final application

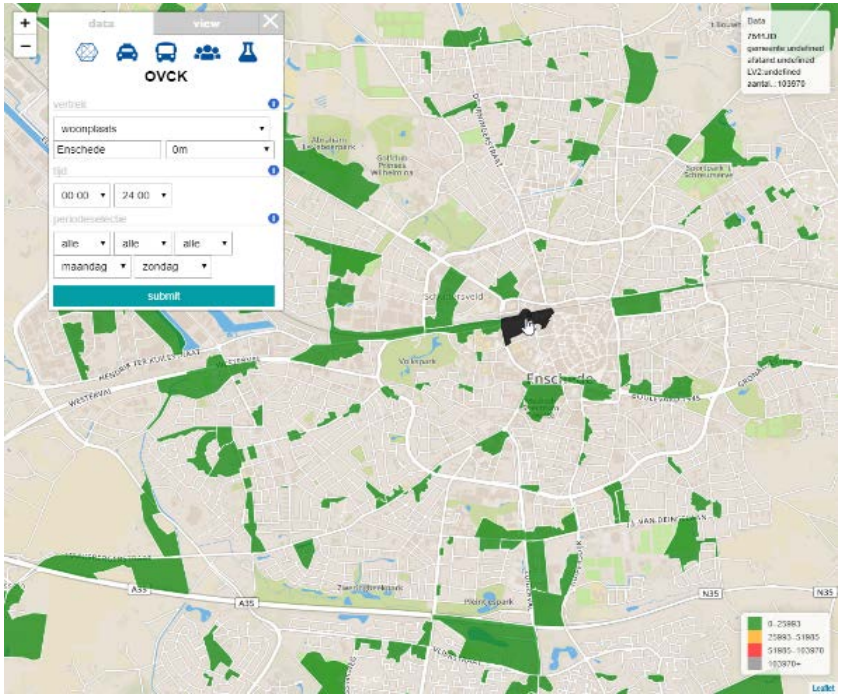


*Figure 15: Final concept application overview*

The final application of which Figure 15 is a screenshot, has an interface which consists of a selection form which is also the navigation in the top left corner, dynamic information section in the top right corner and a legend in the bottom right. These blocks are positioned on top of the map which shows dynamic information in polygons, which being queried based on the selections made in the selection form.

### *Navigation*

The most important navigation element is the selection form. This form consists of a few icons in the top representing different types of data which can be shown on the map. Selecting one of the icons changes the choices below. The taxi screen for example gives the possibility to differentiate between WMO or OV trips, and has the possibility to exclude trips more than a certain amount of kilometres from the base postal code polygon.

### *Advanced selections*

For more advanced selections it is possible to manipulate the input by using variables in the url bar of the browser. Parameters when available in the selected data which can be manipulated by hand consist of (Table 9);

| Variable | Values | use |
|---|---|---|
| Formname | Ovck<br>Regiotaxi<br>Demography | datasource |
| Pc | ####AA | Postal code origin |
| Distance | # | Destinations within kilometres |
| Coi | # | Circle of influence from origin postcode |
| Data_type | All<br>WMO<br>OV | Selection within regiotaxi data |
| H_to | 0-24 | Hour to |
| H_from | 0-24 | Hour from |
| Dnf | 1-7 | Day name from, Monday (1) till Sunday (7) |
| Dnt | 1-7 | Day name to, Monday (1) till Sunday (7) |
| startgeo | Postcode<br>Woonplaats | Selection within ovck data |

*Table 9: Advanced application parameters*

## Tests, improvements and evaluation of the design

### Database optimization

Testing the first version of one of the queries to show data based on PC4 level on the map had a real bad performance of 23 seconds before the data was loaded on screen. By doing some query-diagnostics it became clear specific indexes where needed to speed up the query and that it was better not to have calculations in the JOIN part of the query. The join originally was a LEFT(pc6,4), where the first four characters of the pc6 column were extracted. To speed this up, the table was updated to have the pc4 added as a new column, however this is not good practice from a theoretical point of view. By also placing an index on the JOIN fields (postcode 4 columns in both tables), the query time was reduced to 0.01 second. A decrease of 99,95% in processing time.

The optimization of queries has been done throughout the whole development. For example the query loading ovck data based on place names was working well, but underperforming with waiting times between 10 and 40 seconds per query. By rewriting the query to use joins instead of sub-queries it now performs in under 0.1 second.

### Testing with colleagues

Showing and testing an early version of the tool raised questions on extra functionality. Based on this feedback a function was added to be able to get insight in the travel relations based on municipality, instead of just postal code with an including distance to the centroid.

It also became clear that it wasn't very intuitive that the size of a polygon had nothing to do with the amount of people traveling there. Extra information needs to be added in the instructions to use this application to explain the results shown. Or in a later version the polygons centroid can be (dynamically) transformed to circles which are bigger when more people travel to a certain location.

| PC | Amount of trips | Location |
|---|---|---|
| 7462MZ | 496 | Verzorgingshuis Maranatha |
| 7461ER | 475 | Stichting Woonvoorzieningen |
| 7461MA | 325 | De Schutse Carintreggeland |
| 7461BJ | 288 | - |
| 7461AG | 285 | - |
| 7463PC | 249 | - |

*Table 10: Most visited location regiotaxi Rijssen*

# Design Deliverables

## Case studies

To show the added value of the tool designed, the DSS was used to try to answer three questions from the daily practice at Keolis;

The manager alternative public transport was curious to see if the newly implemented KeoFlex service in Rijssen was competing with the public transport already servicing this town.

From the same department the question was asked what the success rate would be for the new neighbourhood bus service in Borne.

The works council (ondernemingsraad) was curious to know if there would be potential for a direct bus line between the town of Denekamp and city of Almelo in the morning rush.

These questions are used in case studies as a means to test the usefulness of the tool in practice to answer questions from daily practice at Keolis.

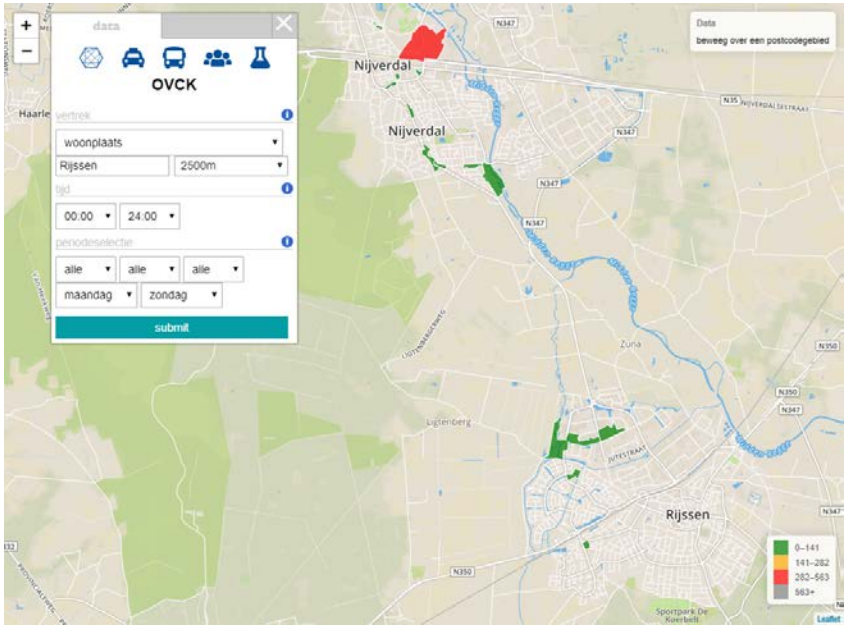


*Figure 16: OVCK use with origin Rijssen*

### Case Study: Rijssen

July 2018 Keolis launched a new service (KeoFlex) in the city of Rijssen to replace the regiotaxi service. The service, KeoFlex, can be characterized as a taxi service with limited origins and destinations with the possibility for the operator to combine trips.

To see if this service competes with the regular bus, the DSS can be used to conduct an analysis on current patterns of bus users in the city of Rijssen (Figure 16) and compare this with the target audience of the KeoFlex service which can be found in the regiotaxi data. As can be seen in Figure 16 there is almost no use of the bus within the town of Rijssen. People tend to travel mostly towards Nijverdal Station,

probably to get into the train in the direction of Zwolle, as this will be a quicker / cheaper route than going by train to Almelo and switch trains there.

To see the potential and areas of interest for the new service, a map was generated using the DSS (Figure 17). In which the centre of Rijssen was chosen as centre point, combined with a origin of 2500 meters and destination of 2500 meters. In short, this gives all trips made within a circle of 2500 meters around the centre of Rijssen. The red areas are the areas to which the most trips are made. The KeoFlex service can expect to get most reservations from these areas (Table 10).

Concluding this case study; the new KeoFlex service doesn't seem to be a threat to the existing bus lines servicing Rijssen, as most trips made now go to outside the city. Based on trips made by the regiotaxi service, most potential can be seen near elderly homes. Based on these results it is advisable to advertise in these areas and focus planning in combining trips in these areas.

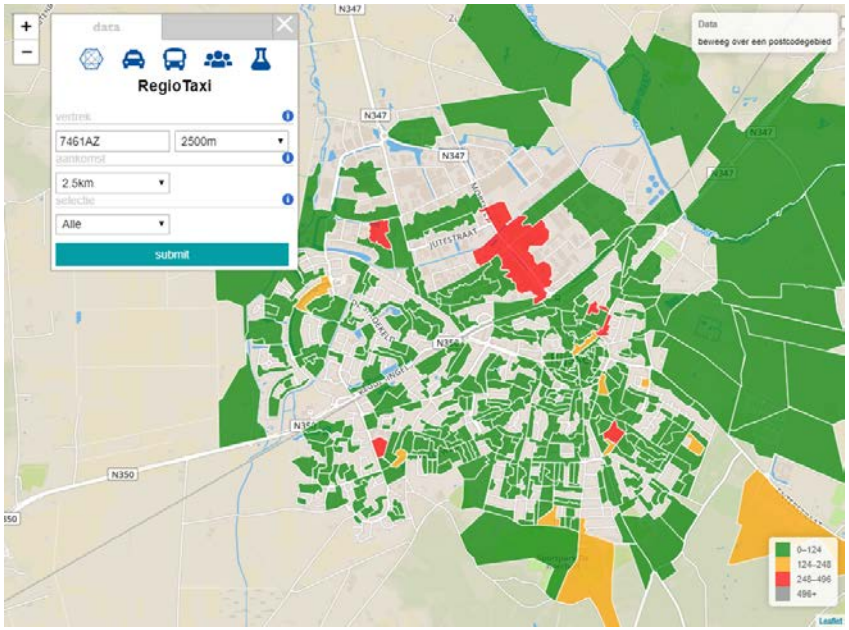


*Figure 17: Regiotaxi trips 2500m circle centre Rijssen*

## Case Study: Borne

In Borne line 30 recently was replaced by a neighbourhood bus. This bus, driven by voluntary drivers, has three routes of about 20 minutes which it alternately drives within one hour. In this case we want to examine the spatial distribution of public transport users. Based on data from OVCK (Figure 18) and regiotaxi (Figure 19) it will be possible to determine what can be expected of the new neighbourhood bus service.

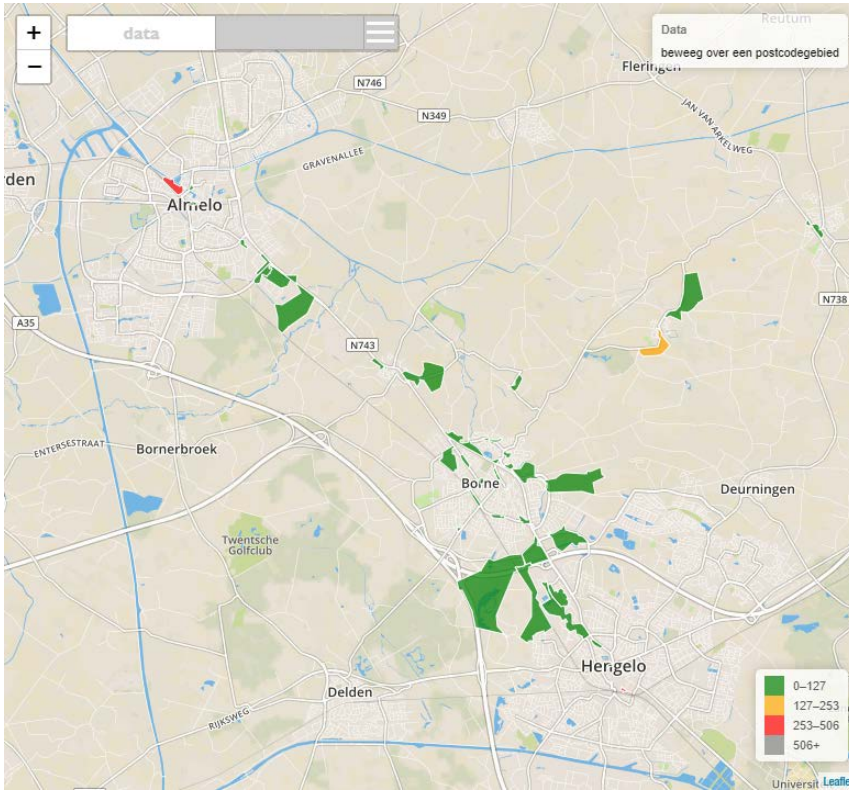


*Figure 18: OVCK use with origin Borne*

Based on the OVCK data it can be concluded that people from Borne mostly use the bus to get to Hengelo and Almelo central station. A few trips are made in the town itself, to the station (125 trips in march 2018) and to IKEA (98 trips in march 2018). Based on OVCK data there seems to be little potential for a service within the town of Borne.

When the regiotaxi data is being analysed, it becomes clear most trips are, just as with the bus, made to locations outside of the town. Most interaction can be seen

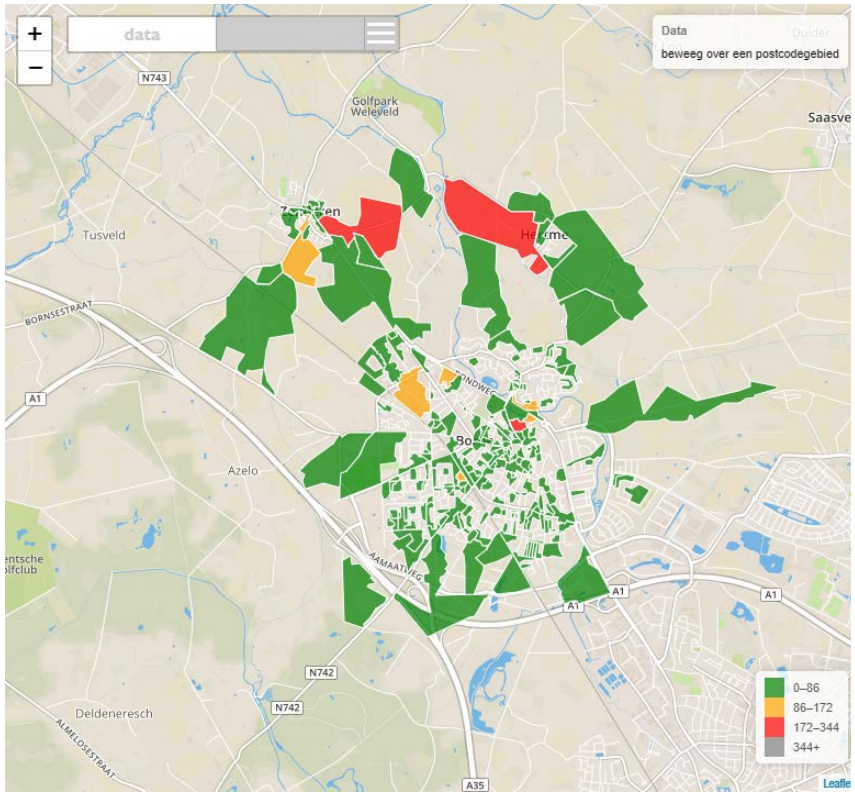


*Figure 19: Regiotaxi trips 2500 circle centre Borne*

between Borne and the neighbouring small townships of Zenderen and Hertme. Only one zone stands out (7622CM), an elderly home (Het Dijkhuis) in Borne. A few people do travel from within Borne to this location.

Based on the data available, OVCK and regiotaxi, it seems there is not much need for public transport within the town of Borne. Most trips made by bus and regiotaxi have a destination outside of the town. This can be explained by the fact that Borne is a small town in which most regular services can be reached within walking or short cycling distance. The farthest distance from the centre (city hall) to the outskirts of town is 2km walking distance.

## Case Study: Denekamp - Almelo

The works council (ondernemingsraad) Twente wanted to know if there would be enough potential patrons to justify a bus service between Denekamp (+/- 9.000 inhabitants) and Almelo in the morning rush as this would be an upgrade in comfort for the clients not having to transfer in Oldenzaal. Logic would dictate there would be quite some potential for a direct line between these municipalities. Almelo, as well as Hengelo, is home for a big regional vocational school (MBO) to which quite a large part of 16-20 year olds have their daily education. This is a group of public transport captives as the distance is too far to use a (regular) bike. Most of them are not allowed to drive a car themselves, and at least a large part of this group of students doesn't have the financial space to afford a licence and costs involving the use of a moped or scooter, let alone a car.

The assumption is people traveling between Denekamp and Almelo now use the bus to Oldenzaal and from there switch buses as this is the fastest (1:01 hour) and cheapest (€7,54) option. Using journey data, data which connects trips into a journey when the time between check-out and check-in is less than 35 minutes, Figure 20 shows the OD-map based on bus travel with Denekamp as a base.

The result is quite disappointing, as only 17 journeys could be identified in a whole month (working days between 6 and 9 in March 2018). Which means on average only 1 person a day uses the bus to get from Denekamp to Almelo.

As quite a substantial part of the users are public transport captives, which means they have to use PT as a means to get to school, they have to be using the train, which is not available in the current data. Although it is slower (1:04), more expensive (€7,73) and an extra transfer is necessary, more people seem to be using this route. This can be explained by different causes;

Since January 2017 traveling by PT is also free of charge for all people going to the MBO. This means the cost of the trip is not important anymore in the choice of a

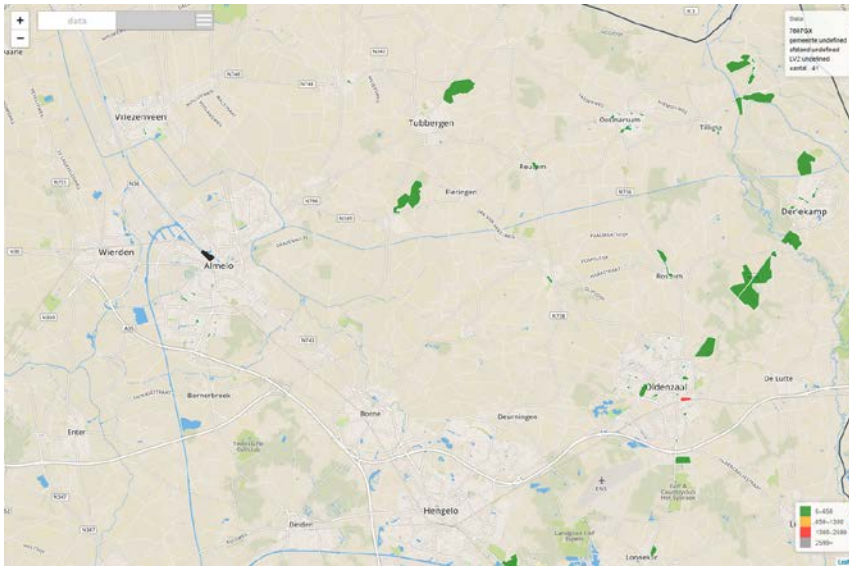


*Figure 20: OVCK journeys Denekamp - Almelo*

route. As travel time only differs 3 minutes between a bus or a double train transfer the travel time is also no big hindrance. Even less so as the train also has a more reliable service, as it is not hindered by rush hour traffic, and the transfer in Hengelo has 6 connections per hour to Almelo in case the connection fails within 10 minutes there is another option. Next to that the service level of a train is higher than the service level of a bus, moreover as the connecting bus travel has a transfer at a regular bus stop (Vos de Waelstraat) which is not comfortable at all in case of wind and rain.

Based on these premises the conclusion is drawn that probably the large majority therefor chooses to travel by train from Oldenzaal to Almelo, of which the data wasn't

available in the data warehouse at the moment which means no potential could be estimated for a fast morning rush connection between Denekamp and Almelo.

A method of researching the total potential would have been to use data from before January 2017, as kids which had not turned 18 yet didn't have a free PT-card. Therefor the price of the trip was far more influencing the choice of route. As it was impossible to buy a subscription including the train between Hengelo and Almelo, kids were more or less forced to use the bus to bus transfer in Oldenzaal. This historical data would have been a great source of information, but the subscriptions were on paper, so no digital record was made.

Further investigation into travel patterns between these two municipalities therefor should consisted of adding train data to the journey generation process. As data of the train between Oldenzaal and Hengelo is available as it is driven by Keolis under the Blauwnet label, this was added. Also starting 2018 two of the six trains an hour traveling between Hengelo and Almelo are Keolis driven and therefor data is available which can be matched to the bus data by identification number of the OVCK and a finding a connecting trip within 35 minutes, the same criteria as used by a connecting bus trip.

Using this enhanced journey data unfortunately the amount of people traveling from Denekamp to Almelo doesn't increase that much to only 45 in a months' time. 2 per working day is still a lot less then would be suspected based on demographics and signals from professionals in the field.

Explanations for this unexpected result, and therefor a reason to dive even deeper into the data, could be the quality of the connection between the Blauwnet trains, as it is possible to use trains from the Dutch railways on this journey as well. If this is the case this would lead to an over representation of people checking out in Hengelo, as this is the station where people need to transfer trains and they end their Keolis travel there. To make this estimation more precise also the returning journeys should be included into the estimation, as someone who returns daily from Almelo to Denekamp but doesn't travel towards Almelo every morning is probably using the NS trains in the morning.

Unfortunately there is also no over representation of people checking out in Hengelo, which leads to believe there could be another factor in play. A possible explanation could be bad behaviour by the users of a student card. As they travel 'for free' it could be they don't take the check in and check out procedure seriously and therefore just don't end up in the data. In the bus this process is regulated more or less, as the driver works as a means to enforce the right behaviour. On a station no one checks if the user is properly checked in or out, the station is not closed by entrance gates. This in mind could mean an over representation of people checking out at station Oldenzaal. Which is indeed the case. To verify this premise some field work will be necessary; observing the check in behaviour of people at the station of Oldenzaal.

Keep in mind while estimating the potential between Denekamp and Almelo, also the improvement in service level should be taken into account in the success rate of a direct connection. The travel time could become considerably lower, no connections have to be made and the route is shorter than the current PT options, all factors making the service more reliable. This could lead to people now using a moped or scooter to also consider the bus.

All in all this case study shows the prototype does what it is supposed to, support the decision making process. It points out a counter intuitive result, which forces the user to look deeper into the problem to find a better explanation for a phenomenon.

## Prototype description

As can be seen in the case studies, the prototype is working and usable in the analysis of spatial and temporal public transport patterns. As it is a prototype a lot can be improved still. The possible improvements will be discussed in the section on future work which will be treated later on.

## Techno-economic feasibility

The tool and underlying methods developed can have quite some added value for Keolis. Several colleagues have indicated they see it as a positive innovation which gives them more insight in information they can use in their daily work. In its current state, as being a prototype, the tool isn't ready to use in production though. Quite some usability bugs are still in the software; the database is static and therefor not updated with new data automatically and static data cannot easily be added yet. All these functions can be implemented in developing a stable version of the tool.

For this tool to be stable enough for company use, further development is necessary. At the moment the focus, and in turn budget, of Keolis isn't on quantitative analysis of data. This means potential cost reduction and potential extra income which can be found using data analysis is left untapped as the available data isn't used to its full potential. Potential in relation to extra income is mentioned here, as it is quite hard to quantify on beforehand what the value of a BI analysis will be, compared to the relative easy sum of, among other, saved diesel and bus driver salary when a bus can be removed from the time table. Direct financial gain therefor is hard to determine from further developing this tool and methods which give more insight in travel patterns and behaviour of (potential) customers.

As the focus and investments at the moment are not in favour for developing this tool further the feasibility is at the moment deemed low. This however could change quickly if budget will be made available as the tool is relatively cheap to maintain and develop further whereas gains could be big with a better insight in the data available.

## Impact

During development data is used which is only available for research. The polygons used are a paid resource available for the university only. Removing them could in theory seriously impact the functionality of the tool. Fortunately also polygons were found from an open internet source, making it possible to still use the tool. These polygons weren't tested as thoroughly as the ones used during development though. Probably the impact on usability of the tool as of deletion of the licenced polygons will be low.

Another means of 'impact' which is of importance, is when it is looked at from the viewpoint of what the impact of the tool and process will be now it is finished. As described in the previous paragraph, the focus within Keolis isn't on quantitative data analysis. The impact of the tool developed depends heavily on the way the board of the company treats the transition to more digitalization and quantitative data research. With investments in data analysis a real impact in knowing the customers and current company processes can be made by making use of and developing for example this tool. Impact in work flows and knowledge obtained by making use of data analysis could be high, but as the focus within Keolis momentarily is on current operations and in the foreseeable future no significant investments will be made in data analysis, the impact of this specific tool will be low.

# Conclusion and Future work

For this conclusion it is good to return to the objectives of this PDEng project;

- Travel and demographical data identification, appraisal and preparation
- Data mining in order to create OD-matrices
- Designing an easy to use tool to display travel patterns in Twente

In short, these objectives have been met in the project. Data from different sources was obtained, appraised and when found useful prepared and put into a data warehouse. For OVCK and regiotaxi reservation data O-D matrices can be constructed for use in the tool that displays travel patterns of an area on postal code 6 level.

During the project techniques were discovered to prepare and analyse the data, these techniques proved helpful in gaining more insight in the data and offered possibilities to do other things with the data than was planned on beforehand. These analyses have also been treated in this report.

The final tool is functioning. It shows patterns based on historical OVCK and regiotaxi data. Selections can be made on dates an time slots. Using this tool it is possible to do case studies answering practical questions. For Rijssen it seems the elderly / assisted living homes are locations with a highly concentrated user base. For Borne it is clear the people living here tend to move outside of the town, which makes it less interesting to provide an inner town public transport service. For the case Denekamp - Almelo the tool shows interpretation is a very important part of using this tool. If only the data shown is used without background knowledge, the connection between these two municipalities doesn't seem to exist, but with some logic and reasoning the conclusion is that more research is necessary to come to a proper conclusion.

The tool is proof of concept in that it is possible to construct visualizations of historical travel data. As with a lot of projects, during the process more and more applications become clear which are interesting to incorporate into the, in this case, decision support tool. Out of time constraints which mostly happened because of a planning which was a bit optimistic, there was too little time in the end to incorporate more functions into the tool. What still would be nice additions to the DSS or studies to do with the data available:

## Tool additions: design related

Additions to the functionality of the tool which are most important if developed further and are more or less production or design related consist of:

Bug fixing; The tool is functional, but is still riddled with functionality bugs for which time within the PDEng programme was too short to fix them all. The first step into developing the tool into a production worthy tool would be to fix the most apparent bugs.

Export functions; At the moment a screenshot has to be made to export data from the tool. It would be good to provide a clean screen without interface which can be easily exported as an image. Another development could be to offer an export with tabular data.

The tool can be expanded to use live data from the production database. Mostly backend automation has to be added to make this possible as some aggregating tables will have to be generated and updated automatically on a regular interval.

Different outputs should be created as well. Being able to output data on different scales could be of interest. For example using PC4 polygons or a grid instead of the current PC6 polygons. Or for example using symbols / dots increasing in size instead of using coloured PC polygons.

## Tool additions: Data analysis

Extra functionality on the analysis side of the tool, which are more research related consist of;

Prediction value; At the moment the tool shows travel data based on factual trips. This means latent potential of areas which are not serviced by PT yet, are underrepresented. By using regression or machine learning techniques based on properties of an area an estimation can be calculated of the untapped potential if a PT-service with a certain level of service will be introduced. In other words, trip generation.

Extra data; More data sources can be added to build a more complete information profile on the areas in the tool. For example the influence of weather on patronage could be added, or later on when available, OD's on a low level of scale from mobile phone location data. Data on black patronage could be added to assist in assigning inspectors to the bus routes which have the highest risk profile. Also data on the KPI's punctuality and customer satisfaction can be added to the tool to give it more functionality and a broader use case for people working at Keolis.

# Literature


Alexander, I. (2005). A Taxonomy of Stakeholders: Human Roles in System Development. *International Journal of Technology and Human Interaction, 1*(1), 37.

Alsger, A., Mesbah, M., Ferreira, L., & Safi, H. (2015). *Public transport origin-destination estimation using smart card fare data.* Paper presented at the Transportation Research Board 94th Annual Meeting.

Appleton, B. (1997). A Software Design Specification Template. Retrieved from http://www.bradapp.com/docs/sdd.html

Bostock, M. (2012). Line Simplification. Retrieved from https://bost.ocks.org/mike/simplify/

Clahsen, A. (2017). Hoe Connexxion zich opnieuw uitvindt. *Financieel Dagblad*.

ESRI. (2018). Grenzen en plaatsen. Retrieved from http://www.esri.nl/producten/content/content/grenzen-plaatsen

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From Data Mining to Knowledge Discovery in Databases. *AI Magazine, 17*(3).

Geofabrik. (2018). Geofabrik.de. Retrieved from https://www.geofabrik.de/data/download.html

Gkiotsalitis, K. S., Antony. (in prep). *Predicting the Traveling Distances and Unveiling the Mobility/Activity Patterns of Individuals from Multi-source Data*.

Groningen, R. (2016). Postalcode 6 areas Netherlands. Retrieved from http://opendata.rug.nl/datasets/ee6772a214fc4c13b8cd2993516e4417_1

Han, J., Pei, J., & Kamber, M. (2011). *Data mining: concepts and techniques*: Elsevier.

Hofmann, M., Wilson, S. P., & White, P. (2009). *Automated identification of linked trips at trip level using electronic fare collection data*. Retrieved from

Hollevoet, J., De Witte, A., & Macharis, C. (2011). Improving insight in modal choice determinants: an approach towards more sustainable transport. *Urban Transport XVII: Urban Transport and the Environment in the 21st Century, 116*, 129.

Imergis. (2017). Geografische open-data GIS bestanden. Retrieved from http://www.imergis.nl/asp/47.asp

Janssen, J. (2017). *TRP: Mobility innovations in Twente: Possibilities to bundle RegioTaxi trips into small-scale concrete public transport connections*. Retrieved from

Kepinski, W. (2018). Markt voor mobiele diensten krimpt verder in Nederland. *Dutch IT Channel*. Retrieved from https://dutchitchannel.nl/602379/markt-voor-mobiele-diensten-krimpt-verder-in-nederland.html

Korotayev, A. V., & Tsirel, S. V. (2010). A spectral analysis of world GDP dynamics: Kondratieff waves, Kuznets swings, Juglar and Kitchin cycles in global economic development, and the 2008–2009 economic crisis. *Structure and Dynamics, 4*(1).

Kurauchi, F., & Schmöcker, J.-D. (2017). *Public Transport Planning with Smart Card Data*: CRC Press.

Li, T., Sun, D., Jing, P., & Yang, K. (2018). Smart Card Data Mining of Public Transport Destination: A Literature Review. *Information, 9*(18).

LISA. (2018). Retrieved from https://www.lisa.nl/home

Malley, B., Ramazzotti, D., & Wu, J. T. (2016). Data Pre-processing *Secondary Analysis of Electronic Health Records*: Springer, Cham.

Mapbox. (2018). Retrieved from https://www.mapbox.com/

Menken, I. (2013). *Data Mining Complete Certification Kit - Core Series for IT*: Emereo Publishing.

OV-Chipkaart. (2017). Overstappen. Retrieved from https://www.ov-chipkaart.nl/zo-werkt-reizen/hoe-werkt-het-reizen/overstappen.htm

Pelletier, M.-P., Trépanier, M., & Morency, C. (2011). Smart card data use in public transit: A literature review. *Transportation Research Part C: Emerging Technologies, 19*(4), 557-568. doi:https://doi.org/10.1016/j.trc.2010.12.003

Robusto, C. C. (1957). The cosine-haversine formula. *The American Mathematical Monthly, 64*(1), 38-40.

Rogers, E. (2003). *Diffusion of Innovation*: Simon and Schuster.

Schmeink, B. (2018) *CEO Transdev: "We zitten allen gevangen in verouderde bedrijfsmodellen"/Interviewer: H. Middelweerd*.

Steenbrugge, X., & Dedecker, M. (2015). *Analysis of personal and company-based carpooling incentives in Belgium.* (MSc.), Universiteit Gent, Gent.

Stewart Fotheringham, A., & Rogerson, P. A. (1993). GIS and spatial analytical problems. *International Journal of Geographical Information Science, 7*(1), 3-19.

Telecompaper. (2017). *Smartphone penetration Netherlands 2017 Q1*. Retrieved from https://www.telecompaper.com/account/download.aspx?cid=1195214

Tientrakool, P. H., Y.; Maxemchuk N. (2011). *Highway Capacity Benefits from Using Vehicle-to-Vehicle Communication and Sensors for Collision Avoidance*. Paper presented at the Vehicular Technology Conference (VTC Fall), San Francisco, CA, USA.

Tobler, W. R. (1970). A computer movie simulating urban growth in the Detroit region. *Economic geography, 46*(sup1), 234-240.

UNOOSA. (2018). *World Geodetic System 1984*. Retrieved from http://www.unoosa.org/pdf/icg/2012/template/WGS_84.pdf

van Beijeren, J. H. M., & Dasburg-Tromp, N. (2010). *Taxibranche onderzoek 2009/2010* (313050000). Retrieved from

van Wee, P. (2012). Een paar carpoolers meer veel files minder. Retrieved from http://www.volgensnederland.nl/themas/mobiliteit/een-paar-carpoolers-meer-veel-files-minder

Wp 2000, (2000, 6 juli).

Wu, B., Sekely, P., & Knoblock, C. (2012). *Learning data transformation rules through examples: preliminary results.* Paper presented at the IIWeb '12, Scottsdale, Arizona.