Postmasters study program

Mathematics for Industry

Social Self-Sufficiency Predictive Model as Decision Support for Local Policy Makers

Final Report

Prepared by
E.K. Sergidou

Prepared for
GGD Hart voor Brabant
GGD West Brabant

31 January 2017

ISBN-INFORMATION

Eleni-Konstantina Sergidou

Eindhoven University of Technology 2017.
Final reports of the Professional Doctorate in Engineering (PDEng) program:
Mathematics for Industry
ISBN [number]
Keywords: self-sufficiency GGD, ROCK, prediction, random forest

ISBN [number]

Social Self-Sufficiency Predictive Model as Decision Support for Local Policy Makers

Eleni-Konstantina Sergidou

Mathematics for Industry
Department of Mathematics and Computer Science
Eindhoven University of Technology, the Netherlands
P.O.box 513, 5600 MB, Eindhoven

University supervisors: Stef van Eijndhoven

Eindhoven University of Technology

Industrial supervisors: Lian Schaerlaeckens

Wieteke de Vries

GGD Hart voor Brabant

Ike Kroesbergen

Judith Helmink

GGD West-Brabant

ABSTRACT

The municipalities in The Netherlands are responsible for funding and providing appropriate social care to their residents by implementing the Social Support Act (Wet Maatschappelijke Ondersteuning, WMO). The GGDs assist their municipalities in developing a local policy for social care according to the needs of the residents of those municipalities, for the coming years. The goal of this project was to predict the self-sufficiency in each district in the municipalities associated with the GGDs, HvB, and WB, with a prediction horizon of four years. Using the GGD data, we could estimate the self-sufficiency levels of adults per district for the years 2005, 2009 and 2012. We researched on the predictability of the self-sufficiency level using external data sources, such as CBS. The research did not show a high correlation the generated self-sufficiency data and the CBS data. For the self-sufficiency rank of a district, the CBS data gives a strong prediction. The procedures developed during this project have high generic quality. They can be applied on the elders in the Monitor data and on data for 2016. We recommend using our scheme and developing it further, by looking for other data sources such as the municipalities.

EXECUTIVE SUMMARY

Since the beginning of 2015, by law, the municipalities in The Netherlands have been responsible for funding and providing appropriate social care to their residents by implementing the Social Support Act (Wet Maatschappelijke Ondersteuning, WMO). The GGDs have to assist their municipalities in developing a local policy for social care in accordance with the needs of the residents of those municipalities, for the coming years. The goal of this project was to predict the self-sufficiency level for each district in the municipalities associated with the GGDs, HvB, and WB, with a prediction horizon of four years.

To reach this objective, we use data (called Monitor data) collected from questionnaires that the GGDs sent to well sampled populations of the 45 municipalities in the years 2005, 2009, and 2012. In addition, we use open data provided by Statistics Netherlands (Het Centraal Bureau voor de Statistiek, CBS).

From the available datasets (besides the Monitor and CBS data), we lack historical data on the total number of self-sufficient people per four digit postcode (PC4) or district. Since we have no data of the quantity we are interested in, we decided to use the Monitor data to estimate it for the years 2005, 2009, and 2012. From the Monitor data, we can derive the quantity of interest at PC4 level. Therefore, we developed a procedure that allows us to extract a feature (e.g., the number of observations) per district from the Monitor data.

We worked on the definition of self-sufficiency and we decided to use the Self-Sufficiency Matrix (SSM) developed by the GGD Amsterdam as our guide. Our SSM divides a person’s daily life into nine domains, such as Finances, Day-Time Activities (DTA), and Mental Health. In each domain, a person can be characterized from self-sufficient to non-self-sufficient. We decided to use three or four categories of self-sufficiency depending on the domain and the available data.

For each domain, we clustered (grouped) the adults (19 to 64 years old) in the Monitor data based on the level of similarity of their answers to questions relevant to the domain. A three-person team from TU/e and the GGDs, HvB and WB categorized each cluster based on the

pattern of their answers. For the years 2005, 2009, and 2012, we derived the percentage of adults in each self-sufficiency category per district and for all the municipalities together.

Our next step was to investigate which factors influence the percentage of people in each category. Therefore, we connected our results with the CBS data about district characteristics, such as average property value and population density. Based on our analysis, we concluded that the predictive power of the CBS data for the percentage of self-sufficient adults per domain per district is rather low.

We split the districts to top and low rated districts according to their percentage of self-sufficient adults. For every domain in our SSM, we constructed an explanatory model that helps us report the most important factors inside the CBS data and their impact on the self-sufficiency rank of a district. For the Finance, DTA, and Mental Health domain, we were able to use the CBS data to build a predictive model with sufficient quality to predict the self-sufficiency rank of a district, for the years 2013 to 2016 (since 2012 is the last year we can estimate the quantity of interest). A visualization of the results is available. Along the way, we were able to predict the CBS data used in our prediction models for the years 2013 to 2016.

With the prediction on the self-sufficiency rank of a district and the identification of influencing factors, the policy advisors of the GGDs can assist the policy makers of the municipalities in their tasks of arranging social care for the residents of their municipalities.

The final step is the development of a software tool that allows the GGDs, HvB and WB, to upload data that they gather in the year 2016 and obtain predictions for the years 2017 to 2020. The software tool will be developed by trainees of the PDEng program Data Science. With this tool, the GGDs will be able to apply the same procedure to the elders (65 years old or older) in the Monitor data. The clustering technique used in this project can be applied independently from the prediction model, and so the GGDs, HvB and WB, can use it in other projects where appropriate. Furthermore, we suggest considering data from other data sources to improve the prediction models developed in this project.

CONTENTS

1 MANAGEMENT INTRODUCTION

1.1 PROBLEM BACKGROUND

The authorities in The Netherlands want to stimulate the Dutch citizens to take responsibility for solving their social-care-related problems within their possibilities using their social networks. Whenever this is not possible, support can be provided by volunteers and/or professionals in the social care system. Since the beginning of 2015, by law, the municipalities in The Netherlands have been made responsible for funding and providing appropriate social care to their residents by implementing the Social Support Act (Wet Maatschappelijke Ondersteuning, WMO).

The Community Health Services (Gemeentelijke Gezondheidsdienst, GGD) monitors the health and welfare of the residents in the municipalities assigned to it. The GGDs have to assist their municipalities in developing a local policy for social care in accordance with the needs of the residents of those municipalities, for the coming years. To define a policy that can cover the demand for social care, preferably at low cost, it is crucial to predict the self-sufficiency levels of the residents.

The problem background is based on Arum (2015); Arum and Schoorl (2016); Klerk, Gilsing, and Timmermans (2010); Kluft and Vlaar (2015); Meinema (2014); Peelen, Anniek, Holland, and Exter, (2010).

1.2 PROJECT GOAL

The GGD Hart voor Brabant (HvB) in ‘s-Hertogenbosch and the GGD West-Brabant (WB) in Breda are responsible for the public health of citizens in their regions, which consist of 27 and 18 municipalities, respectively (1,742,134 people in total). The goal is to predict the self-sufficiency level of each district in all the municipalities associated with the GGDs, HvB and WB, with a prediction horizon of four years.

Besides predicting the self-sufficiency level of a district the two GGDs want to know which factors influence this level. Having this knowledge, the policy advisors of the GGD can assist the policy makers of the municipalities in their tasks of arranging social care for the residents of their municipalities.

1.3 AVAILABLE DATA

Data collected from questionnaires that the GGD sent in the previous years is provided and is used for this purpose. This data, called Monitor data, is at an individual level and it covers a wide spectrum of a person's daily life.

Apart from the Monitor data, other data sources, accessible to the GGDs HvB and WB, should have also been taken into account. A list of available data sources was provided by both GGDs. The greater part of the open data is available at a municipality or district level. In most cases, data at district level is available only for a few years. Therefore, for this project, we considered open data from Statistics Netherlands (Het Centraal Bureau voor de Statistiek, CBS).

1.4 PROJECT PROCESS

We began the project on 1 January 2016 and we finished it on 31 December 2016. We divided the project in three phases. In this section, we provide a summary of the work done and our results in each phase.

Phase 0 – Literature Review and Data Exploration: We focused on the understanding and description of the problem definition, which led to the problem background and goal as presented in this report. We studied and explored the available data to identify their assets and shortcomings. The Monitor data is split per age group (adults and elders) and per year (2005, 2009, and 2012). This division occurs by the different questionnaires sent in each age group and year. For all the individuals in the Monitor data, we know their four-digit postcode area (PC4), but not their district.

The CBS data, as other open data, is available at district level, providing information, such as average property value, population density, and average income. Furthermore, we lack historical data on the total number of self-sufficient people per PC4 or district. Since we have no data of the quantity we are interested in, we decided to use the Monitor data to estimate it for the years 2005, 2009, and 2012.

We agreed on a definition of the concept of self-sufficiency that we used during this project. Our definition is based on a simplified version of the Self-Sufficiency Matrix (SSM) developed by the GGD Amsterdam (Lauriks, et al., 2013). The original SSM consists of 11 domains referring to characteristics of a person's daily life, e.g., Finances, Day-Time Activities (DTA), and Mental Health. We excluded two domains of the SSM (Housing and Judicial) because no data is available for these domains.

In each domain, a person can be characterized from self-sufficient to non-self-sufficient. The original SSM proposes five categories of self-sufficiency. We decided to use three or four depending on the domain and the available data. The new categories are the results of merging two self-sufficiency categories of the original SSM.

Phase 1 – Estimation of the self-sufficiency levels per district: We decided to focus on the adults in the Monitor data, since they form the backbone of the society, and thus, play an important role in realizing a participation society. However, at all time, we kept our procedures as generic as possible in order to be applicable to the elders as well.

We developed a method that allows us to extract a feature of interest (such as number of observations) from the GGD data at a district level. We checked how many observations with complete information per district are within the Monitor data in order to determine if we can reach conclusions related to the whole population of adults.

Using a part of the data of the GGDs, HvB and WB, from 2012, we categorized the adults according to their level of self-sufficiency applying two approaches. Based on an analysis of the advantages and disadvantages of each method, we selected ROCK (Guha, et al., 2000), a clustering technique, as our method to estimate the self-sufficiency levels per district.

We continued by applying this technique to the adults in the Monitor data. For each domain, we clustered (grouped) the adults (19 to 64 years old) in the Monitor data based on the level of similarity of their answers to questions relevant to the domain. We (a three-person team from the TU/e and the GGDs HvB and WB) categorized each cluster based on a pattern of their answers. At this point, information related to background characteristics, such as ethnicity and place of residence, was not taken into consideration. We aimed to prevent coloring our judgement and instead to label the individuals based only on their ability to take care of themselves. For the years 2005, 2009, and 2012, we derived the percentage of people in each self-sufficiency category per district and for all the municipalities together.

Phase 2 – Predictive and explanatory models: In this phase, we visualized the trend of the self-sufficiency level of the adults through the years. We observed that, for the Finance, DTA, and Addiction domains, there is an upward trend, while, the self-sufficiency level in Physical Health domain decreases from 2009 to 2012. For the Mental Health and Domestic Relation domains, self-sufficiency level remains approximately the same.

Furthermore, we connected our results on self-sufficiency with the CBS data. From early results in this phase, we concluded that we cannot accurately predict the percentage of self-sufficient adults for a district. The reason is that the CBS data describe a district at a higher level than required in order to predict the self-sufficiency level of the adults. Therefore, we split the districts to top and low rated districts according to their percentage of self-sufficient adults. For the Finance, DTA, and Mental Health domains, we used the CBS data to build a random forest model with sufficient quality to predict the self-sufficiency rank of a district, for the years 2013 to 2016. At that stage, we considered 2013 – 2016 as our future, since 2012 is the last year for which we know the self-sufficiency level of the districts.

However, having in mind that the GGDs, HvB and WB, are gathering data from 2016, they need to be able to follow the same procedure and forecast 2017 to 2020. For this task, forecasts of the CBS data are required. Therefore, using linear regression, we developed a prediction model for each predictor variable in the CBS data.

Besides the obtained predictions for the self-sufficiency level of the adults in a domain, the GGDs, HvB and WB, are interested in identifying which factors influence this level. We performed this identification task by visual means and by constructing an explanatory model for every domain in our SSM. For these models, we used the data from 2012. For most of the domains, it appears that the average house value is an important factor.

1.5 Conclusions and Recommendations

One of our main conclusions is that, based on the Monitor data, we could estimate the self-sufficiency levels of adults per district for nine domains. We researched on the predictability of the self-sufficiency levels using external data sources, such as CBS. The research did not show high correlation for all the domains. For the self-sufficiency rank of a district, the CBS data has strong predictive quality for three out of the nine domains.

The procedures developed during this project have high generic quality. Therefore, they can be applied on the elders in the Monitor data and on data for 2016, as well. We recommend using our scheme and developing it further, by looking for other data sources such as the one that can be provided by the municipalities associated to the GGDs, HvB and WB.

2 PHASE 0: LITERATURE REVIEW AND DATA EXPLORATION

Phase 0 was dedicated to understanding the problem contexts. Besides a literature review on social care in The Netherlands, we studied available data and we worked on a self-sufficiency definition. In the following sections, we provide an overview of our understanding and findings related to these topics.

2.1 GGD DATA: MONITOR DATA

In the years 2005, 2009, and 2012, each GGD sent questionnaires to the residents of its associated municipalities. The GGDs divide the residents into two groups:

1. Adults (people who are 19 years old or older and younger than 65)
2. Elders[1] (people who are 65 years old or older).

The items included in these questionnaires cover a wide spectrum of fields such as physical health, welfare, environmental concern, and housing. The questionnaires of different years and age groups are not identical, since questions were added, changed, or removed. Each GGD gathered the completed questionnaires into datasets and processed the responses of all the questionnaires in order to create reports on the health and well-being of the residents of each municipality. Different datasets correspond to different age groups and years, because of the different questionnaires sent. The total of datasets is called Monitor data. Despite their differences, all datasets share common characteristics.

The Monitor data is at an individual level and is stored in SPSS files, where each row (observation) corresponds to one completed questionnaire. The variables in each dataset correspond to the items of the questionnaire and to their outcome after being processed by the GGD. With few exceptions, the variables in the Monitor data are categorical. We have access to the raw and the processed data. We decided to work with the processed dataset, since it is a clean version of the raw dataset. For each person who completed a questionnaire, we know his/hers municipality and living area (defined by a four-digit postcode, called PC4). For all the individuals in the dataset, their self-sufficiency level is unknown.

[1] The GGD sent questionnaires to the elders in 2006.

2.2 OPEN DATA SOURCES

As mentioned in the introduction, other data sources, accessible to the GGDs, HvB and WB, were also taken into account. An overview of the various data sources was provided by the GGDs. Based on this overview, we discovered that the largest amount of open data is available at a municipality or district level. Moreover, we could not find the total number of self-sufficient people in a data source, regardless of the area used (district or PC4).

We studied open data from Statistics Netherlands (*Het Centraal Bureau voor de Statistiek*, CBS). This government agency is one of the largest open data sources in The Netherlands, concentrating on demographic, social, environmental, and economic data. In the CBS data[2], one can find information related to population, houses, and income for each district of a municipality, for example.

Later in this project, a trainee from the PDEng Data Science, Sarah Ibrahimi, explored in depth various data sources from the list given by the GGDs. Her goal was to investigate what data is available in those sources and whether it is of use in our project. She created an overview, where for each data source, she explains the topic of the data, at which level it is available, the years of collection, and the way to extract the data. Her main conclusion was that, indeed, CBS offered the best data at that point of time.

2.3 SELF-SUFFICIENCY DEFINITION

Parallel to studying available data, we worked on the definition of self-sufficiency that we used during this project. Starting from the new concept of positive health (Huber, 2014), we went to the Self-Sufficiency Matrix (SSM) created by the GGD Amsterdam (Lauriks, et al., 2013). The main reason was that the positive health concept concentrates on the overall health and abilities of an individual. In contrast to that, the SSM has a broader reach and includes domains of a person’s life, such as financial situation and domestic relations.

The SSM contains 11 domains, namely Finances, Day-Time Activities (DTA), Housing, Domestic Relations, Mental Health, Physical Health, Addiction, Activities of Daily Life (ADL),

[2] http://statline.cbs.nl/Statweb/

Social Network, Community Participation, and Judicial. The SSM is made to categorize an individual from completely self-sufficient to not self-sufficient at all and is based on an extensive interview with the individual. In Table 1, we give a brief explanation of the self-sufficiency categories in the SSM.

We simplified the SSM, since we had to categorize the individuals based on their answers provided by the questionnaires instead of from a personal interview. Thus, instead of having the five categories as there are in the original SSM, we merged two or more categories into one for each domain. The last two rows in Table 1 show the categories used in this project. In addition, there was no data in the Monitor data that could help us categorize the people in the domains Housing and Judicial. Therefore, we ignored these two domains.

Table 1: Self-sufficiency categories in the SSM

Acute problems	**Not self-sufficient**	**Barely self-sufficient**	**Adequately self-sufficient**	**Completely self-sufficient**
The situation is untenable. There are acute problems.	The individual is not self-sufficient. Situation will deteriorate if there is no intervention.	The individual has limited self-sufficiency. The situation is stable, but barely adequate.	The individual is adequately self-sufficient.	Self-sufficiency is above average.
Non-self-sufficient		**Barely self-sufficient**	**Adequately self-sufficient**	**Self-sufficient**
Non-self-sufficient		**Almost self-sufficient**		**Self-sufficient**

2.4 Results

From the available datasets, we lack historical data on the total number of self-sufficient people per PC4 or district. The Monitor data provides information of a person in various aspect of his/her life. Therefore, having defined self-sufficiency, we decided to use the Monitor data to estimate the percentage of people in each self-sufficiency category in the years 2005, 2009, and 2012. The number of categories used depends on our data.

We chose the dataset of adults for 2012 (Adults 2012) as our starting point. We first applied and evaluated a possibly appropriate method on that dataset before we employed the method to the rest of the datasets. We decided to start with the adults in the Monitor data, since they form the backbone of the society, and thus, play an important role in realizing a participation society. At all time, we kept our procedures as generic as possible in order to be applicable to the elders, as well.

We worked at district level, since most of the open data sources, including the CBS, have more data at district level than at PC4 level.

In the next section, we explain how we proceeded with our work and how we derived the percentages of adults in every self-sufficiency category from the Monitor data.

3 PHASE 1: ESTIMATION OF THE SELF-SUFFICIENCY LEVELS PER DISTRICT

As mentioned in the previous section, for the years 2005, 2009, and 2012, we decided to estimate the percentage of people in each self-sufficiency category using the Monitor data. For four domains of the SSM, we categorized the adults in the Monitor data from 2012 according to their level of self-sufficiency using two approaches. Based on the advantages and disadvantages of each method, we selected one of them to be applied to the rest of the domains. Throughout this project, we used the municipalities of Bergen op Zoom and ‘s-Hertogenbosch to demonstrate and check our results.

From the Monitor data, we could calculate the desired percentages per PC4. There is more open data at district level than at PC4 level. Therefore, in Phase 1, we developed a method that assigns a district to a PC4 area. This method allows us to extract a feature of interest from the Monitor data at district level.

3.1 VARIABLE SELECTION

Our first step to estimate the percentage of adults in each self-sufficiency category was to choose the variables we to be used in each domain. This process started in February and we reached at the final set of variables per domain in July 2016.

For each domain of the SSM, we selected the variables from the Monitor data that can help us to determine the self-sufficiency level of a person in this domain. Thus, we choose variables that indicate whether a person is self-sufficient. For example, for the Finance domain, we chose the variables that indicate if the individual: 1. earns an above average annual income, and 2. has a difficulty with paying the bills.

To avoid coloring our judgment, we excluded variables, such as ethnicity and living place, since factors like those may explain the self-sufficiency of a person, but they do not determine whether that person is self-sufficient or not.

As mentioned in Section 2.1, the questionnaires sent throughout the years are not identical. Some questions have been altered, removed, or added. Hence, not all the variables are available in every year.

Since we started with the data from 2012, we used our choice of variables from Adults 2012 as our guide for 2009 and 2005. We searched in Adults 2005 and 2009 for the variables that correspond to those from 2012. In case there was no variable that corresponds to a variable from 2012, we searched for one with a similar meaning. We could not find a match for every variable. This implies that there is not sufficient data for the ADL and Social Network domains in 2005 and 2009. In addition, in 2005, we have no data for the Community Participation domain.

3.2 Logic Model vs ROCK

We used two approaches applied to Adults 2012 to categorize adults according to their level of self-sufficiency in the domains of Finances, Day-Time Activities, Domestic Relations, and Mental Health. Our goal was to choose a suitable approach for our data contexts and datasets. The two methods were

- **Logic Model**: A data specific technique. Using the variables that describe a domain of the Self-Sufficiency Matrix, a three-person group from the GGDs HvB, and WB and Eindhoven University of Technology determined the rules for someone to be Self-sufficient, Almost self-sufficient, or Non-self-sufficient. The logic model was based on the concept of decision tree. Hence, for each domain, we weighed the variables according to their definition. Based on the value of each variable, starting from the most important one, we determined the level of a person's self-sufficiency. Each entry in Adults 2012 is labeled according to these rules.

- **ROCK**: **A robust clustering algorithm for categorical attributes** (Guha, et al., 2000): A data driven technique. For each domain, the entries in Adults 2012 were clustered (grouped) based on their similarities in the variables describing the domain. An important parameter of ROCK is $theta$. With this parameter, the user of ROCK defines how much similar two entries/clusters must be in order to be merged to a higher order cluster ($theta$ takes values from 0 – not similar – to 1 – the same). The algorithm stops when there are no lower order clusters that can be merged into a higher order one. Based on the profile of each cluster, I labeled it Self-sufficient, Almost self-sufficient, or Non-self-sufficient.

Regardless of the method, we calculated the percentage of adults in each category for all the domains. Figure 1 gives these percentages out of all individuals in Adults 2012 per domain.

To compare the two methods, we used the same variables and the same number of self-sufficiency categories for both. The restriction came from the logic model, since we had to explicitly predefine which variables and categories are used in order to build our logic model. Using ROCK, though, we obtained an overview of our data, since we saw how many clusters and what kind of clusters there are. That led to a better understanding of our dataset before we labeled the clusters in comparison to the logic model. In other words, after the application of ROCK and based on its results, we could discuss and decide which self-sufficiency categories are appropriate for a domain. In addition, we could explore which variables we have to include or exclude from our analysis. Because of the aforementioned reasons, we decided to use ROCK as our clustering technique.

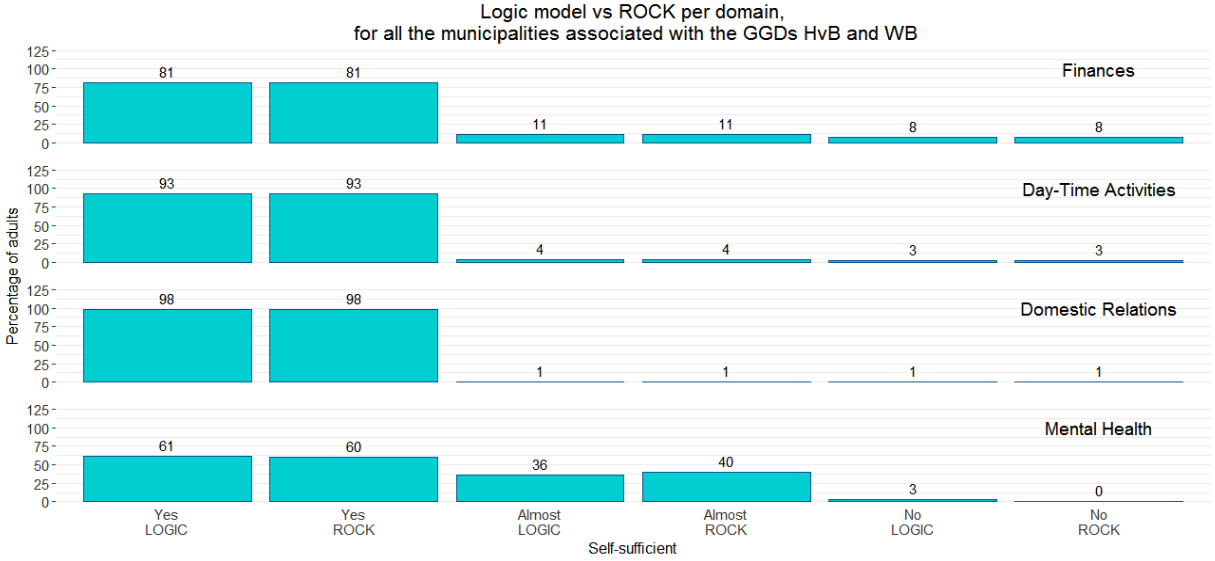

Figure 1: Percentage of adults in each self-sufficiency category given by the Logic Model and ROCK. The results are for all the municipalities and for the domains of Finances, Day-Time Activities, Domestic Relations, and Mental Health.

3.3 APPLYING ROCK ON ADULTS

Choosing ROCK as our clustering technique, we proceeded with applying this technique on the dataset Adults. For every domain and year, the individuals in Adults were clustered based on their responses in the variables describing the domain.

To show the clusters generated by ROCK, we plotted their profiles. Thus, for each cluster, we could see what kind of answers the people in this cluster gave. Using these profiles, a three person group from the GGDs HvB and WB and Eindhoven University of Technology interpreted the results and labelled the clusters.

After discussions about the interpretation and labelling, the choices of variables used in a domain were evaluated and updated if needed. A new choice of variables for a domain led to re-applying ROCK on Adults 2012 and re-interpreting the results. For each domain, we repeated this process at least three times before we made our final choice of variables used in every domain.

As mentioned earlier, the user of ROCK has to choose a value of $theta$, the similarity parameter for allowing two entries/clusters to be merged. For all domains and years, by trial and error, we chose those $theta$'s that provided the clearest image of the groups inside Adults, and thus, enabled an easy interpretation of the results.

Using the method described so far, all adults in the Monitor data were assigned to a self-sufficiency category per domain. This results in the self-sufficiency levels of all 45 municipalities together. Since we could not compare these results with actual data, we validated our method using background characteristics of the individuals (as mentioned in Section 3.1, we excluded variables such as age and ethnicity from our analysis so far). We visualized the relationship between the self-sufficiency categories and the background characteristics: age, ethnicity and education level. We compared the expectations based on the experience of the GGDs, HvB and WB, with our visualization of 2012. Since we met the expectations, we consider our method as being validated.

3.4 POSTCODES (PC4) AND DISTRICTS

As mentioned in Section 2.1, for each person in the Monitor data, we know his/hers PC4 but not the district. Since we worked at district level, we had to figure out the relation between districts and PC4s. For this purpose, we used data from CBS, where, for each address in The Netherlands, its municipality, district, neighborhood, and PC4 are given. This dataset is updated by the CBS every year. For our analysis, we used the one from 2015. Additionally, we restricted ourselves to the municipalities associated with the GGDs, HvB and WB. Using this data, for each PC4, we found how many districts it belongs to and what these districts are. We calculated the percentage of the PC4s that belong to each of these possible districts. Figure 2 shows the relationship between the PC4s and the districts in ‘s-Hertogenbosch. As we see, there are PC4s that belong to more than one district.

Thus, we developed a procedure that we apply every time we want to calculate a feature per district, such as the number of observations, from the Monitor data. According to this procedure, if 95% or more data of a PC4 belongs to one district, we assign this PC4 to that district only. In any other case, we randomly choose a district using the percentage of the PC4 in this district as the probability of belonging to that district. We calculate the value of the desired feature per district and repeat the process as many times as wanted. In the end, we calculate the average value of the desired feature per district. Moreover, we show how the PC4s are assigned to districts for the municipalities of Bergen op Zoom and ‘s-Hertogenbosch.

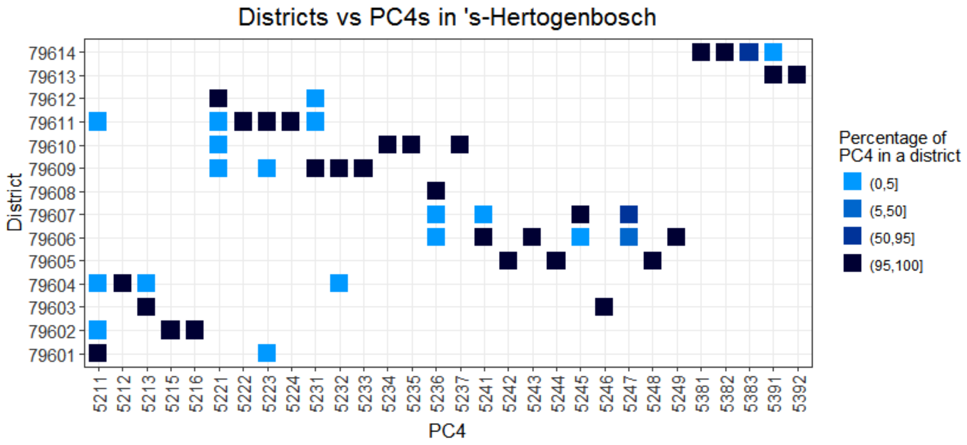

Figure 2: Relation between PC4s and districts in ’s-Hertogenbosch

3.5 Observations with Complete Information

Complete cases are observations for which the values of all variables are known. For each year, the number of complete cases per district is a feature of Adults that we were interested in, because we wanted to know the size of our sample and whether we could reach conclusions related to the whole population. As mentioned earlier, for each domain, we chose various variables that can help us determine the self-sufficiency level of an individual in Adults. In other words, for each year, the dataset of adults is divided into subsets, each of which contains some of the variables in Adults.

For 2012, Figure 3 shows the percentage of complete cases per domain in Bergen op Zoom, ‘s-Hertogenbosch, and for all the municipalities. Additionally, it provides the percentage of complete cases when we combine all the domains. In Appendix M, we provide the results for each district in Bergen op Zoom and ‘s-Hertogenbosch for the years 2005, 2009, and 2012.

As shown, for each domain, the percentage of complete cases is at least 75% (the minimum is given by the Physical domain). When we combine the domains, the percentage drops to 60%. We note that, for all the domains and for both municipalities, approximately 80% of the observations in each district is complete. Therefore, we decided to take only complete cases per domain into account. If there are fewer than 50 complete cases in a district, we ignored this district from our analysis. Table 2 gives the number of districts with 50 or more complete observations per domain for the years 2005, 2009, and 2012. The last column provides the number of districts with 50 or more complete observations in every year.

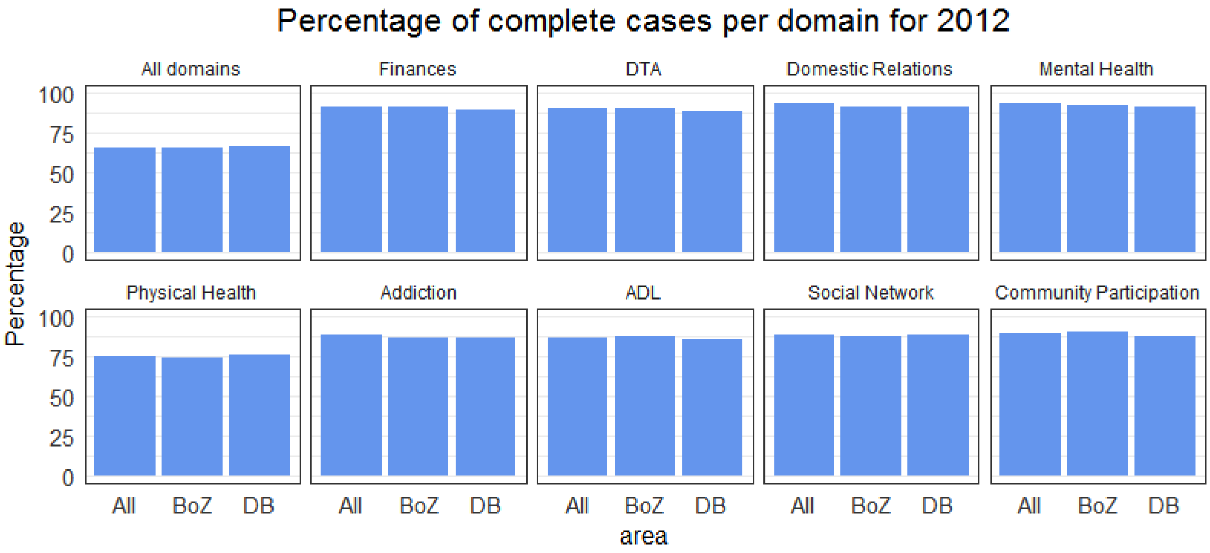

Figure 3: The total number of observations in all municipalities (All), Bergen op Zoom (BoZ), and 's-Hertogenbosch (DB) is 24641, 710, and 1497, respectively. The plot shows the percentage of observations that are complete per area per domain and for all the domains together.

Table 2: Number of districts with 50 or more complete cases per domain per year and in all years

Domains	2012	2009	2005	all
Finances	137	99	106	82
Day-Time Activities	137	124	127	103
Domestic Relations	137	127	133	109
Mental Health	137	125	133	106
Physical Health	115	115	125	90
Addiction	133	123	122	99
ADL	134	-	-	134
Social Network	136	-	-	136
Community Participation	135	122	-	110

3.6 Results

In Phase 1, we focused on estimating the percentages of adults in each self-sufficiency category per domain using the Monitor data. Our first step was to choose those variables inside the data that could help us determine a person's self-sufficiency level. We concluded that, for the ADL and Social Network domains, we could describe the self-sufficiency situation of the adults only for the year 2012, while we could for the Community Participation domain, for the years 2009 and 2012.

To derive the percentages of interest, we applied two approaches (logic model and ROCK) on a part of the Monitor data. The main difference of the two approaches is that, in the logic model, we first determined the logic rules that decide whether an individual is self-sufficient and the number of categories that we should have. In contrast, by applying ROCK, we let our data speak and then we decided what label to assign to each cluster. Therefore, we used ROCK for the rest of the Monitor data.

For the years 2005, 2009, and 2012, using ROCK, each entry in Adults was assigned to a cluster. After labelling the clusters, we derived the percentage of people in Adults per self-sufficiency category, for every domain. In addition, we developed a procedure that allows us to derive a feature of interest from the Monitor data at a district level. Hence, we could derive the percentages of adults in a self-sufficiency category per district. For a district, we accepted the derived percentages only if they were based on at least 50 observations with complete information. As an example, Figure 4 shows the percentages of adults in each self-sufficiency category for the Finance domain in Bergen op Zoom in 2012.

As mentioned above, we developed a procedure that allows us to derive a feature of interest from the Monitor data at a district level. We note that this procedure is generic and so can be applied outside the scope of this project. A similar observation holds for ROCK. We used ROCK as an intermediate step in our analysis, however, ROCK can be used in research, and where clustering of data with categorical attributes is required.

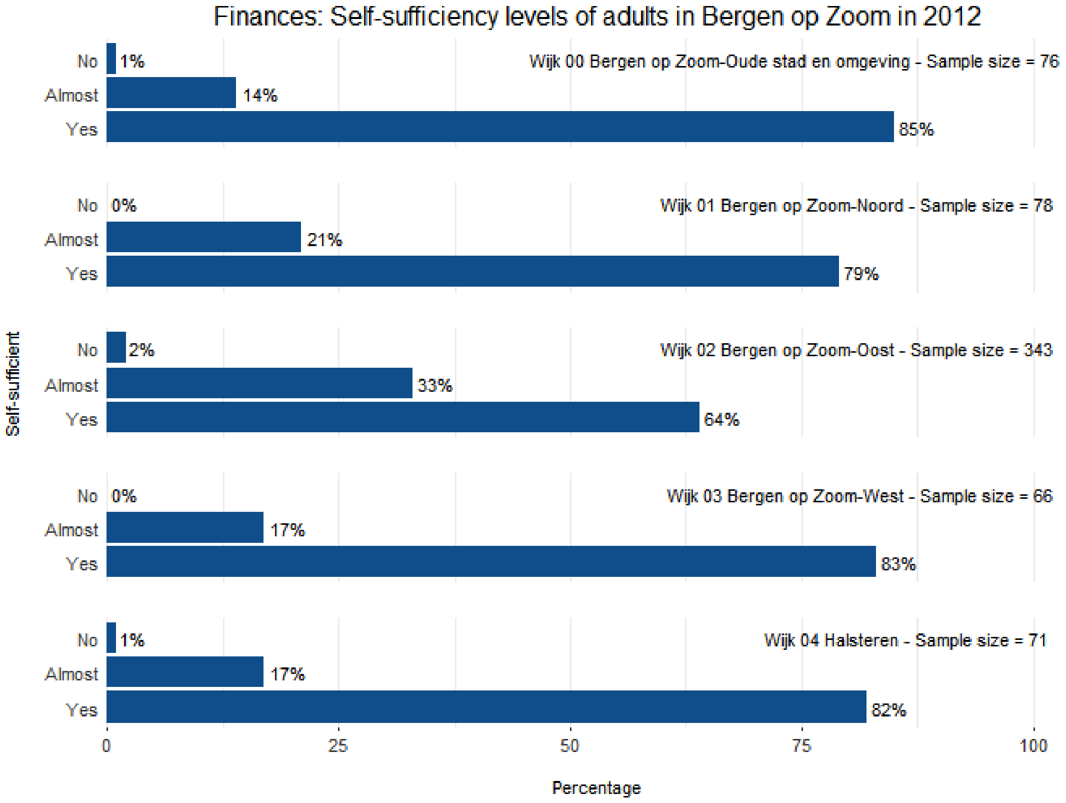

Figure 4: Self-sufficiency levels of adults in Bergen op Zoom in 2012 for the Finance domain

4 PHASE 2: PREDICTIVE AND EXPLANATORY MODELS

Having the results from the previous phase, we continued by exploring the trend of the self-sufficiency level of the adults through the years, as well as, the relationship between the domains. One of our main goals in Phase 2 was to develop a model to predict the percentage of adults in each category per domain. To achieve this goal, we connected our results on self-sufficiency with CBS data that helps us characterize a district, such as average house value, population density, and average income.

At that stage, we considered 2013 – 2016 as our four-year horizon, since 2012 is the last year we knew the self-sufficiency level of the districts. Since the GGDs, HvB and WB, are gathering data from 2016, they need to be able to follow the same procedure to forecast 2017 to 2020. Therefore, besides predicting the self-sufficiency levels of a district, forecasts of the CBS data are required too. For that, we developed a prediction model for each predictor variable in the CBS data.

From preliminary results, we concluded that we cannot accurately predict the percentage of self-sufficient adults for a district using the CBS data. Therefore, we focused on predicting the rank of a district (high or low percentage). For the Finance, DTA, and Mental Health domain, we used the CBS data to build a model with sufficient quality to predict the self-sufficiency rank of a district, for the years 2013 to 2016.

Besides predicting the self-sufficiency rank of a district, the GGDs, HvB and WB, were interested in identifying which factors influence this rank. We performed this identification task by visual means and by constructing an explanatory model for every domain in our SSM. For these models, we used the data from 2012.

4.1 SELF-SUFFICIENCY OF ADULTS THROUGH THE YEARS

Since we were interested in forecasting self-sufficiency, we explored how the percentage of self-sufficient adults changes from 2005 to 2012. We focused on the six domains for which we estimated this percentage for 2005, 2009, and 2012. In other words, we explored the change through the years for the Finance, DTA, Domestic Relation, Mental Health, Physical Health and Addiction.

Using boxplots and histograms as our visual means, we were able to obtain an overall view of the trend in the domains mentioned above. We observed that for the Finance, DTA, and Addiction domains, there is an upward trend, while, the self-sufficiency level in the Physical Health domain decreases from 2009 to 2012. For the Mental Health and Domestic Relation domains, self-sufficiency levels remain approximately the same.

4.2 Domains' Relationship

Parallel to the work described in the previous section, we explored the relationship between the domains in the SSM. To achieve this goal, we compared the percentage of self-sufficient adults per district for every pair of domains using scatterplots and correlation matrices. We noticed that, mainly, there is a positive but weak correlation between each two domains.

To have a better understanding of the results, we thought about the self-sufficiency at an individual level. As in the case of the background characteristics, being self-sufficient in Physical Health, for example, does not necessarily imply a high self-sufficiency level in Finances. In other words, the self-sufficiency level in one domain may affect the level in another domain, but it does not define it.

4.3 Processing and Exploring the CBS Data

In this project, we used the CBS data to predict the self-sufficiency in a district. For every year between 2005 and 2016, we downloaded data about the neighborhoods, districts and municipalities of the GDDs, HvB and WB, from the official website of the CBS.

In each year, though, the data is available for the districts and municipalities existing in that year. For example, the district Vinkel belonged to Maasdonk until 1-1-2015, when this municipality ceased to exist. Now, it is a district in the municipality of ‘s-Hertogenbosch. So, for the years 2005-2014, one can find data about this district by searching in Maasdonk, and for later years in ‘s-Hertogenbosch.

From the beginning of our project, we have used the districts as they were in 2015. Hence, we made sure that the CBS data referred to these districts by following the procedure below.

Compare the districts existing in a year with the ones existing in 2015. If the code of that district is the same, assume that there is no significant change to the district. Otherwise, old and new districts are compared and their relation is determined, which is one of the following:

1. One neighborhood becomes a district. The data of the neighborhood is used as the data of the new district.
2. Two or more old districts or neighborhoods are merged into one new district. In this case the data of the old districts/neighborhoods are simply combined to derive data of the new district.
3. Two or more old neighborhoods are merged with parts of old neighborhoods to create one new district. In this case first estimate the size of each part in percentage of the split neighborhood and split the data accordingly. Then combine the (split) data of the old neighborhoods to derive data of the new district.

Appendix R provides all changes in districts and municipalities from 2005 to 2015. In addition, we provide examples of calculating the data of a district when merging of splitting occurs.

For each year, we performed an exploration of the CBS data by checking the correlation of each variable with the other variables available for that year. We noticed high correlations between some of the variables, for example, between the percentages of owned and rental houses. Using both of the highly correlated variables in a prediction model leads to redundancy, since they provide the same information. Therefore, we processed those variables first, so that they reflect the situation in the district more adequately than the unprocessed ones. For example, we calculated the ratio of the owned houses to the rental houses. All the correlation matrices are given in Appendix S.

4.4 Exploring the Relationship Between the CBS Data and the Self-Sufficiency Level of Adults

By employing visual means, we explored the relationship between the CBS data and the percentage of self-sufficient adults in a district. We created correlation matrices for the years 2005, 2009, and 2012. Moreover, for these years, we created scatterplots for each domain and each variable in the CBS data (predictor). These correlation matrices and scatterplots can be found in Appendix T.

In each year, we performed this exploration for all possible domains. In other words, we provided scatterplots for the Social Network domain for 2012 only, since we have no estimation of self-sufficiency level of the districts for 2005 and 2009.

We observed low correlations between the CBS data and the percentage of self-sufficient adults. In some scatterplots we observed a trend, but in most cases, a trend was difficult to distinguish.

Besides visual exploration, we built a regression model per domain for 2012. In other words, we fitted a linear model using the data of 2012. The results showed that the CBS data can explain only a small portion of the variation in the percentage of the self-sufficient adults, to be more precise: 30-50%. The Social Network domain is an exception, since the CBS data explains 70% of the variation for this domain.

4.5 Predicting Self-Sufficiency Rank

We decided that we do not use the self-sufficiency data of earlier years as an input, so that our model is also applicable for districts for which we could not estimate its self-sufficiency in one of those years. The model thus only depends on the CBS data of that year. Subsequently we could build one model for 2005, one for 2009, and one for 2012 and compare them.

Moreover, based on the results discussed in Section 4.4, we decided not to predict the percentage of self-sufficient adults. We transformed our numerical variable (percentage of self-sufficient adults) into a categorical variable. To be more precise, we decided to predict the rank of the districts. The rank of a district in a year is calculated by comparing its percentage of self-sufficient adults in that year with the median percentage of all districts. If the percentage of a district is equal to or higher than the median, the district is ranked as “1” (in top 50% of the districts). Otherwise, the district is ranked as “2” (in bottom 50% of the districts).

Additionally, since we wanted that a model built for one year is also applicable to other years, we transformed the CBS data into categorical data. For the CBS data we used four categories, the numerical values were split using the four quantiles. We note that we only used variables from the CBS data that were available for all years since 2005.

We decided to use a random forest as our prediction model. A random forest is a big collection of decision trees. Each tree is built using approximately 70% of the observations and a random subset of the predictors of a user-defined size. Each tree is validated by the 30% that was not used to build it. The accuracy of the forest is the average of the accuracies of the trees. To predict the self-sufficiency rank of a new observation, all decision trees are applied. The resulting ratio forms the prediction, e.g. if 56% of the decision trees decide that the new observation has rank " 1", then the outcome of the random forest model is the prediction that the new observation has rank "1".

For the years 2005, 2009, and 2012, we constructed a random forest per domain to predict the rank of a district based on the predictors from the CBS data. We observed similar accuracies of the models of different years for each domain. For the six domains we had self-sufficiency data for all these years, we tested the models of each year on the data of the other years. Once again, the accuracy in other years was similar to the accuracy for its own year. Therefor we chose the models of 2012 as the final ones. Based on the accuracy of the random forests we accepted the models for the Finance (68%), DTA (69%), and Mental Health (65%) domains.

A trainee of the PDEng program Data Science, Adam Zika, worked on the visualization of these results. Using R and shiny, he provided a graphical user interface where one can see on a map whether a district is rank "1" or "2" for a domain in a given year.

4.6 Explanatory Models for Self-Sufficiency Rank

Besides the obtained predictions for the self-sufficiency rank of the districts in a domain, the GGDs, HvB and WB, are interested in identifying which factors in the CBS data influence this rank. We performed this identification task by constructing an explanatory model for every domain in our SSM. For these models, we used the data from 2012, the latest year for which we estimated the self-sufficiency rank of the districts. Moreover, we used every available variable in the CBS data that year. The idea was that we do not restrict ourselves on data that have measurements in all the years, but investigate among every possible variable in the CBS data.

Using random forests as in the previous section, we derived which variables in the CBS data contributed the most in the accuracy of the models, as well as, which variables can distinguish the districts with rank “1” from those with rank “2” better than others. Besides a random forest, we constructed a decision tree per domain, which we plotted to visualize the result. From the decision tree, one can observe easily which values of a variable in the CBS data corresponds to rank “1” or “2” districts.

4.7 PREDICTING THE CBS DATA

As mentioned throughout this report, we considered the years 2013 to 2016 as our four year horizon to predict self-sufficiency rank. Meanwhile, the GGDs, HvB and WB, are gathering the 2016 data and they need to be able to follow the same steps to forecast 2017 to 2020. For those years, of course, we do not have any CBS data; hence forecasts of the CBS data are required. We developed a prediction model for each predictor variable in the CBS data using linear regression.

For each predictor variable, we created a dataset where each row corresponds to a district and each column to a year starting from 2005 until 2012. Then, using the 2012 values as output and the values in the previous years as input, we established a linear relationship between the historical values and the expected future value of each CBS variable. Additionally, we explored how many years of history are needed for an accurate prediction. For some predictors we used seven years of history, while for others five.

Having this relationship, we can predict the value of a predictor for the years 2013 to 2016 per district. For example, to predict the average house value in each district in 2013, we use the values in the years 2008 to 2012. For 2014, the prediction of 2013 is used as input together with the values in the years 2009 to 2012.

To test the accuracy of these models, we used the actual data from 2013 to 2016 (note that some variables have data until 2015). As we mentioned in Section 4.5, we categorized the CBS data split using the four quantiles. We performed the same action on the predicted values and we compared them to the actual values. For all the predictors, the best accuracy is given by the year 2013 (the first year from the four year horizon) and, as it is expected, it drops as we predict more into the future. We note that for almost all predictors the accuracy remains above 70%.

4.8 Results

In this phase, we visualized the trend of the self-sufficiency level of the adults through the years. For the Finance, DTA, and Addiction domains, we observed an upward trend, while, for the Physical Health domain an downward trend. For the Mental Health and Domestic Relation domains, self-sufficiency level remains approximately the same. In addition, we explored the relationship of the domains. We concluded that, even though, there is a relationship, it is not so strong to conclude that a domain is redundant. Hence, we built a prediction model per domain.

We connected our results on self-sufficiency with the CBS data. To provide a prediction model with sufficient accuracy, we split the districts to top and low rated districts according to their percentage of self-sufficient adults. For the Finance, DTA, and Mental Health domains, we used the CBS data to build a random forest with sufficient quality to predict the self-sufficiency rank of a district, for the years 2013 to 2016. Along the way, we were able to predict with high accuracy almost every variable used in our prediction models for the years 2013 to 2016.

By visual means and constructing explanatory models, we identified important indicators for a high self-sufficiency rank per domain. Data from 2012 is used for these models, since 2012 is the latest year we have estimation about the self-sufficiency rank of the districts. For most of the domains, it appears that the average house value is a strong self-sufficiency indicator.

At all moments, we kept our procedures as generic as possible. This implies that the same procedures can be applied, as soon as the 2016 Monitor data is collected, to forecast the self-sufficiency rank of the districts for the years 2017 to 2020. Similar observation holds for the elders in the Monitor data.

5 CONCLUSIONS AND RECOMMENDATIONS

During Phase 1, we developed a procedure to estimate the self-sufficiency levels of adults per district using the Monitor data. We provided the policy advisors in the GGDs, HvB and WB, with an overview of the self-sufficiency situation for adults in 2005, 2009, and 2012. Following the same steps, one can apply this procedure on the data collected on 2016 and on the elders in the Monitor data. Hence, the policy advisors in the GGDs, HvB and WB, can have an overview of the self-sufficiency of the elderly people in 2006, 2009, and 2012. As soon as the two GGDs collect their 2016 data, the overview on the self-sufficiency of adults and elders can be expanded to include the situation in 2016, too. Having this overview, the policy advisors know the needs in a district and can compare it with the overview of four years ago. It might be interesting to cluster the districts based on their self-sufficiency levels.

In Phase 2, we focused on the question whether data from external sources, such as CBS, would have indicative quality for the self-sufficiency in a district. Therefore, we switched from percentages in the districts (results from Phase 1) to ranks of the districts (based on underlying statistics). With this switch, we concluded that, for the Finance, DTA, and Mental Health domain, the CBS data gives a high quality prediction according to a random forest model. The random forest models related the self-sufficiency rank estimated by the Monitor data for the year 2012 and 16 attributes of the CBS data used as predictors. We recommend following the same steps when the 2016 data is available both for adults and elders.

Additionally in Phase 2, we built an explanatory model for every domain, which also can be performed using data from elders and for the year 2016. This model provides the policy advisors with variables in the CBS data that are strong indicators for high or low self-sufficiency ranks for each domain.

The procedures developed during this project have a generic quality. At this moment, in order to apply any of these procedures, one should run the R code that corresponds to this procedure. That assumes a certain level of software professionality of the user. Therefore, we recommend developing a software tool that allows the researchers of the GGDs, HvB and WB, to apply any of the procedures developed in this project, even if they lack programming skills.

We recommend continuing this work of predicting self-sufficiency. So far, important steps were made for the GGDs, HvB and WB: self-sufficiency was determined, data science techniques were introduced, and self-sufficiency rank predictive models were built. There is still work to be done: developing the software tool, applying our procedures to elders, and improving the prediction models (either to have better accuracy or to predict percentages instead of rank). We suggest investigating the role of municipality data in predicting self-sufficiency, since the municipalities are expected to have more detailed data about their inhabitants and areas (e.g., about type of houses and number of parks in every neighborhood).

REFERENCES

Arum Silke van and Schoorl Rosanna *Sociale (wijk)teams in beeld [Online] // movisie. - March 2016. - April 4, 2016. - https://www.movisie.nl/publicaties/sociale-wijkteams-beeld.*

Arum Silke van *Social (neighbourhood) teams: must or hype? [Online] // movisie. - October 15, 2015. - May 2, 2016. - https://www.movisie.com/news/social-neighbourhood-teams-must-or-hype.*

Guha Sudipto, Rastogi Rajeev and Shim Kyuseok *ROCK: a robust clustering algorithm for categorical attributes [Journal] // Information Systems Volume 25 Issue 5. - 2000. - pp. 345 - 366.*

Huber Machteld *Towards a new, dynamic concept of health : Its operationalisation and use in public health and healthcare and in evaluating health effects of food [Book]. - Maastricht : Maastricht University, 2014.*

Klerk Mirjam de, Gilsing Rob and Timmermans Joost *The Social Support Act: the story so far [Online] // The Netherlands Institute for Social Research. - March 1, 2010. - May 1, 2016. - https://www.scp.nl/english/Publications/Summaries_by_year/Summaries_2010/The_Social_Support_Act_the_story_so_far.*

Kluft Maaike and Vlaar Paul *New social work in the Netherlands [Online] // movisie. - December 2, 2015. - May 2, 2016. - https://www.movisie.com/news/new-social-work-netherlands.*

Lauriks Steve [et al.] *The Self-Sufficiency Matrix [Online] // selfsufficiencymatrix. - February 2013. - January 2016. - http://www.selfsufficiencymatrix.org/zrm-int.aspx.*

Meinema Thea *New in the Netherlands: social care and support teams [Online] // movisie. - September 22, 2014. - May 2, 2016. - https://www.movisie.com/news/new-netherlands-social-care-and-support-teams.*

Peelen [et al.] *The Social Support Act (WMO) [Online] // Health Policy Monitor. - November 2010. - May 1, 2016. - http://hpm.org/en/Surveys/Erasmus_University_Rotterdam_-_Netherlands/04/The_Social_Support_Act_(WMO).html.*