Master Thesis

Econometrics and Operations Research

Research on the Social Domain

Which demographic characteristics explain specified target groups?

Author:
Ralf Zoetekouw
ralfzoetekouw@hotmail.com

Supervisor:
Dr. Ines Wilms
i.wilms@maastrichtuniversity.nl

Commissioned by:
Thönissen Management en Advies B.V
m.thonissen@thonissen.nl

Maastricht, September 8, 2020

Samenvatting

De verschuiving van de verantwoordelijkheid van het Sociaal Domein van de rijksoverheid naar de gemeenten in 2015 heeft geleid tot financiële tekorten voor bijna alle gemeenten in Nederland. Gemeenten kampen tegenwoordig nog steeds met een grote uitdaging om de decentralisatie succesvol te maken, zowel voor het welzijn van burgers als in financieel opzicht. Om deze uitdaging te voltooien en de integrale werkwijze te verbeteren, moeten gemeenten beter inzicht krijgen in de data van hun voorzieningen. In opdracht van de Arrangementenmonitor van Thönissen Management en Advies B.V is een onderzoek verricht op de data van de voorzieningen Sociaal Domein van 25 gemeenten met 330 wijken en 160.000 cliënten. De studie onderzoekt de dynamiek van huishoudens die ondersteuning nodig hebben en het onderzoekt welke demografische kenmerken de ondersteuning verklaren voor gespecificeerde doelgroepprofielen. Het onderzoek is opgedeeld in drie analyses en uitgevoerd met behulp van de statistische softwaretools R, Tableau en Stata.

Analyse 1: Doelgroepprofielen aan de hand van stapeling van ondersteuning bij huishoudens

De huishoudens met ondersteuning zijn ingedeeld in doelgroepprofielen omdat dit bijdraagt aan de gerichtere, integrale aanpak van gemeente conform de één gezin, één plan-aanpak: alle ondersteuning in een huishouden wordt als één geheel gezien in plaats van dat de ondersteuning per persoon wordt geanalyseerd. De doelgroepprofielen zijn opgebouwd op basis van drie kenmerken:

- Het verschillende aantal soorten ondersteuning (1, 2, 3 of 4 en meer) in één huishouden;
- De leeftijdscategorie van de gezinsleden (jeugd, volwassenen, ouderen);
- Het type ondersteuning (fysiek, mentaal) in één huishouden.

Uit de eerste analyse blijkt dat, met behulp van het K-means algoritme geclusterd op de verschillende soorten ondersteuning, de Silhouette methode en de Adjusted Rand Index, de geconstrueerde doelgroepprofielen statistisch juist zijn gedefinieerd. Dit betekent dat gemeenten de huishoudens voortaan op eenzelfde manier als in Tabel 5 kunnen classificeren zonder dat er bias ontstaat.

Analyse 2: Op- en afschaling van ondersteuning per doelgroepprofiel

Er is bekend hoeveel mensen gebruik maken van voorzieningen per kwartaal, maar het is nog niet inzichtelijk hoeveel mensen er per kwartaal doorstromen naar andere voorzieningen (doelgroepprofielen). Dit is de eerste studie die de dynamiek van de ondersteuning analyseert en daarmee de patronen van de ontwikkeling (wijzigingen) in de ondersteuning van de huishoudens inzichtelijk maakt. Het onderzoek laat zien dat er een relatief hoge toename van voorzieningen is voor huishoudens die geclassificeerd zijn als:

- huishoudens die financiële hulp nodig hebben in de leeftijdscategorie 'volwassenen';
- huishoudens met twee verschillende hoofdarrangementen met de leeftijdscategorie 'jeugd';
- huishoudens met drie verschillende hoofdarrangementen met de leeftijdscategorie 'ouderen'.

De tweede analyse geeft inzicht aan gemeenten welke type doelgroepprofielen in verband met hun 'opschaling' van ondersteuning meer aandacht nodig hebben.

Analyse 3: Relatie demografische buurtprofielen en doelgroepprofielen

De derde analyse onderzoekt welke demografische kenmerken van buurten (CBS data set) de zorg en ondersteuning van de verschillende doelgroepprofielen significant verklaren. Gemeenten kunnen de resultaten van de derde analyse gebruiken voor hun integrale en gerichte buurtaanpak en benchmarking aan de hand van buurtprofielen. Het uitgevoerde onderzoek slaagt erin om de doelgroepen te verklaren op basis van de demografische gegevens op buurtniveau. De resultaten zijn samengevat in Tabel 7 waarin onder andere de volgende interessante significante relaties naar voren komen:

- Hoe hoger het aantal huishoudens met een hoog inkomen in een wijk, hoe groter de kans dat jongeren gebruik maken van de voorziening 'Jeugd-GGZ'.
- Hoe hoger het aantal huishoudens met een laag inkomen in een wijk, hoe groter de kans op ondersteuning Sociaal Domein in deze wijk. Conform verwachtingen betreft dit met name de doelgroepprofielen die financiële ondersteuning (bijstand en bijzondere bijstand) ontvangen.
- In de meeste gevallen hebben doelgroepprofielen meer ondersteuning nodig wanneer het percentage gescheiden personen in een buurt hoger is. De variabele 'gescheiden' bevestigt de hypothese dat jeugdzorg vaker nodig is als de ouders gescheiden zijn: hoe hoger het percentage gescheiden personen in een wijk, hoe hoger het gebruik van jeugdzorg in deze wijk.
- Doelgroepprofielen met de leeftijdscategorie 'ouderen' worden zowel verklaard door hoge als door lage inkomens: onderzoek wijst uit dat de financiële situatie bij ouderen minder belangrijk is dan bij de andere leeftijdscategorieën met betrekking tot het verklaren van ondersteuning.
- Doelgroepprofielen met drie of meer vormen van ondersteuning komen - verdeeld over 330 wijken - te weinig voor om significante conclusies te kunnen trekken over de aanwezigheid van een relatie tussen demografische buurtprofielen en de doelgroepprofielen. Om meer inzicht te krijgen in deze doelgroepprofielen per buurt is nieuw onderzoek nodig.

Abstract

Shifting responsibility of social care from the government to the municipalities has resulted in financial deficits for almost all municipalities in the Netherlands. Nowadays, municipalities are still faced with a major challenge to make the decentralisation successful, both financially and in terms of the well-being of the people. In order to overcome this challenge, municipalities need to have better insight into their social care data in an effort to facilitate and adjust their integral approach and working procedure. This paper provides an understanding of municipalities' social care data by using the Arrangementenmonitor of Thönissen Management en Advies B.V. The study investigates the dynamics of households that need social care and it examines which demographic characteristics explain social care consumption of specified target groups. The research is divided into three analyses and carried out using the statistical software tools R, Tableau and Stata. The first analysis finds that the constructed target groups are statistically well-defined. The reason for constructing target groups is that it contributes to the targeted policy of the municipalities. The second analysis explores the dynamics of the target groups, thus investigating the in-, out- and through-flow. The third analysis examines which demographic characteristics significantly explain the target groups. Municipalities have to choose the targets groups they want to focus on such that they can adjust their policy accordingly. The results of the third analysis can be used to enhance their integral and targeted neighbourhood approach.

Contents

1 Introduction 1

2 Reviewing the Social Domain 2

3 Data 4

3.1 Arrangementenmonitor 5

3.2 Data cleaning 6

3.2.1 Incomplete reporting 6

3.2.2 Age categories 6

3.2.3 Neighbourhoods 7

3.2.4 Providers 7

3.2.5 Construction final data set 7

3.3 Demographic characteristics 8

3.4 Target groups 9

4 Comparing expert-constructed and data-driven target groups 10

4.1 Methodology 11

4.1.1 K-means algorithm 11

4.1.2 Adjusted Rand Index 12

4.2 Empirical results 13

5 Dynamics of the target groups 14

5.1 Methodology: Tableau procedure 15

5.2 Empirical results 15

6 Demographic characteristics of neighbourhoods 19

6.1 Methodology 19

6.1.1 Variable selection 19

6.1.2 Model selection 20

6.1.3 The model 22

6.2 Empirical results 23

7 Conclusion 26

8 Implications future research 27

List of Tables II

List of Figures II

Glossary III

References VII

1 Introduction

In 2015, the Dutch municipalities were confronted with a big change which is known as *the decentralisation.* The government of the Netherlands has decided that three social care facilities should be administratively and financially regulated on a local level. These three facilities are Youth care, Participation and Social Support Act (SSA), resulting in the three new laws Youth law, Participation law and SSA 2015, respectively. These laws are known as *the three decentralisations* and have come into effect as of 2015. The Social Domain is set up in a completely new way. The municipalities, instead of the Dutch government, are now responsible for regulating social care for their citizens. Reasons for the decentralisation are given by the Dutch Parliament (2014) in the Dutch parliamentary papers. These include closer decision making to the people, an integrated approach instead of an individual approach per domain, more responsibility for the people and an attempt to reduce costs of social care facilities (2014a), (2014b), (2014c).

Nevertheless, the decentralisation has resulted in financial deficits for almost all municipalities in the Netherlands. Koster (2019) from the news website *Binnenlands Bestuur* reports that the Dutch government supported 77 municipalities with 200 million dollars at the end of 2018 because of financial deficits. In addition, GGZ Totaal (2020) and Dutch Child Center (2020) report that two thirds of the municipalities have a deficit of 20% and one fifth even has a deficit of more than 40%. The total deficit within social care is around 600 million euros yearly. In an interview with Kieskamp (2019) from the Dutch newspaper *Trouw*, the Dutch minister from public health, Minister De Jonge, argues that it is not only important to financially support the municipalities, but also to address the youth care problem.

Nowadays, municipalities are still faced with a major challenge to make the decentralisation successful, both financially and in terms of the well-being of the people. In order to overcome this challenge, municipalities need better insight into their social care data in an effort to facilitate and adjust their integral approach and working procedure. Thönissen (2016) is convinced that social care domains are interrelated. Therefore, research has to be done on all social care combined within a household, instead of addressing the problem for every domain separately. This is known as the integral approach. Thönissen Management en Advies B.V. invented the Arrangementenmonitor which supports the integral approach. The Arrangementenmonitor gives more insight into the developments of social care facilities for participating municipalities such that the substantive and financial bottlenecks can be tackled.

The purpose of this paper is to investigate the dynamics of households that need social care and to examine which demographic characteristics explain social care consumption of specified target groups. Moreover, this research enables municipalities to gain more insight into their social care data which, in turn, can be used for a targeted neighbourhood approach. This way, municipalities are informed about their social spending decisions and, as an indirect effect, the well-being of their inhabitants will be improved since social care is optimised and only deployed to people that certainly need it. The research is divided into three analyses which are summarised in the following research questions:

- To what extent are the constructed target groups statistically well-defined?
- Which patterns are recognisable in the in-, out- and through-flow of the target groups?

- Which demographic characteristics significantly explain the target groups?

The first analysis classifies households into 27 segmented target groups. Vuik et al. (2016) state that this contributes to a more targeted policy since this provides insight into the differences in support needs of the social domain clients. The second analysis explores the dynamics of the target groups, being one of the first studies to undertake such analysis. This study aims to contribute to this growing area of research by exploring the in-, out- and through-flow for and between target groups. The third analysis investigates which demographic characteristics in a neighbourhood explain social care on household level for every target group. This analysis enhances a targeted neighbourhood approach and investigates whether and which demographic characteristics of a neighbourhood significantly explain social care consumption in households. In addition, municipalities are able to use this as a benchmark. For example, given certain demographic characteristics in a neighbourhood, a certain amount of social care is expected and forecasted.

The paper is organised as follows. Section 2 provides context for this paper by reviewing the Social Domain. Section 3 gives more information about the data that is used in this research and explains the categorisation of social care by Thönissen Management en Advies B.V. The statistical methods and the empirical findings of the first analysis are discussed in Section 4. The second analysis is reviewed in Section 5. Section 6 is concerned with the methodology and empirical results of the third analysis. The results are summarised in Section 7. Finally, in Section 8, some possible limitations of the analyses are reported and it is stated how they could be overcome in future research.

2 Reviewing the Social Domain

Smeets et al. (2020) classify Dutch social care clients into four segmented groups which contributes to the targeted policy of the municipalities. They perform a latent class analysis to identify subgroups for the HNHC population which are high-need, high-cost and chronically ill patients in primary care. They employ a Life Cycle Analysis (LCA) to classify their subgroups. Collins and Lanza (2009) explain the LCA as a person-oriented analysis technique that identifies classes of individuals with similar patterns of personal factors relevant to social care utilisation. Smeets et al. (2020) find that the main differences between the four subgroups are found in demographic and socioeconomic factors. These factors are age, household position and source of income.

Vuik et al. (2016) argue that big data segmentation analysis that divides a patient population into distinct groups contributes to a better targeted approach. After identifying groups that consist of members that share a lot of characteristics inside the group and share less characteristics with members in other groups, the distinct groups can be targeted with social care models and intervention programs tailored to their needs. Motivated by the research of Smeets et al. (2020) and Vuik et al. (2016), this paper constructs target groups as will be explained in subsection 3.4. Furthermore, it examines whether these target groups are statistically well-defined by comparing them to target groups that have been constructed using a clustering algorithm which will be explained in Section 4.

A lot of recent research has been done about the increase in social care in the Netherlands.

KPMG (2020), CBS (2020) and Westra et al. (2018) all report, on average, an increase of youth care and SSA in the Netherlands for almost all municipalities and, on average, a small increase in the other social care domains. However, none of them report whether there are new clients in social care (in-flow), whether these clients do not need social care anymore (out-flow) or whether these clients come from another social care domain (though-flow). The only information that is available for a municipality is that, for example, it has 400 clients of youth care in the first quarter of 2017 and 425 clients of youth care in the second quarter of 2017. Until now, it remains unknown whether there was an increase of 100 clients and a decrease of 75 clients or an increase of 25 clients. Furthermore, it is unknown whether these clients are new clients or come from another social care branch. This paper has great social impact since it investigates, for the first time in literature, the in-, out- and through-flow of clients in social care in Section 5.

Schellingerhout et al. (2020) focus on the sub-domain youth care without residence and investigate the differences between neighbourhoods for this social care service. They found that there are relatively many neighbourhoods in Groningen, Drenthe and parts of Zuid-Holland and Limburg with a high share of users, while in Overijssel and parts of Noord-Holland many neighbourhoods have a relatively low share of users. Furthermore, they investigate which characteristics significantly explain these differences. In their research, they state that the household composition in the neighbourhoods is of great importance for youth assistance without residence. Also, family characteristics (share of assistance users), child characteristics (special education), and environmental characteristics (share of non-western migrants) play an important role. Other characteristics like the distance living from the doctor, divorces, drug use and education level are less related to youth care without a stay. Finally, they investigate whether there are large differences between the actual use and the expected use of youth care without residence, based on the composition and other characteristics of the neighbourhood. They construct a multilevel model that predicts the amount of youth care for most neighbourhoods relatively well. However, in some neighbourhoods the actual usage is much higher or much lower than they would expect, based on the characteristics of the population in those neighbourhoods. The third analysis of this paper, which will be discussed in Section 6, elaborates on the work of Schellingerhout et al. (2020). Motivated by Thönissen (2016), instead of focusing on only one (sub-)domain, this paper conducts research on all domains of the Social Domain since it is assumed that they are interrelated.

In general, not a lot of research has been done, nor is there much data of Dutch social care facilities available. The reason for this is that it is a relatively new field, with the change of responsibility from the national government to the municipalities that had just taken place in 2015. Westra et al. (2018) confirm this. They are one of the first who conducted research for 18 municipalities of Zuid-Limburg. They investigate which demographic and social-economic characteristics explain the youth mental health facilities for different municipalities. They find that there are large differences between municipalities in Zuid-Limburg, both in usage and costs. However, the reason for this is unknown and they conclude that not enough data was available for satisfactory research to draw reliable conclusions. It is important that their research shows that the results between municipalities are not significantly different since there are big differences between neighbourhoods of the same municipality. This way, doing research on a too highly aggregated level (municipality level) gives generalised results. Instead, CBS Statline (2020), Schellingerhout et al. (2020), Batterink et al. (2018) and Engbersen et al. (2018) state that neighbourhoods are of great relevance to the policy of municipalities. Therefore, they declare that research investigating the factors that explain social care should be done on neighbourhood level. Schellingerhout et al. (2020) argue that this is not only because of the fact that most municipalities work with district teams but also because municipalities take specific measures for specific neighbourhoods depending on the situation in those neighbourhoods. As a result, there is a big difference in the amount of social domain facilities between neighbourhoods. This paper is motivated by these findings and uses demographic characteristics on neighbourhood level instead of on municipality level, as will be explained explicitly in subsection 3.3.

At the time of writing, Hameleers and Westra from Maastricht University are doing the same research on household level. The demographic characteristics on household level are secured and because of privacy regulations not available for public research. In their research, they use the same demographic characteristics as used in this research. Therefore, in the near future, it is interesting to compare these findings and to investigate whether the results on household level and on neighbourhood level coincide with each other.

Nijendaal (2014) explains the distribution of money for social care in the Social Domain over the municipalities. It is largely determined by the social and demographic risk profile of the population in municipalities. For example, municipalities with many single-parent families are expected to use more youth care. Furthermore, municipalities with many elderly people are expected to need more household help. However, in their overall report of the Social Domain, Pommer and Boelhouwer (2017) found that these risk profiles only explain part of the regional differences in the use of the three new laws. Even after taking these risk profiles into account, large differences between regions still appear to exist. In a reaction to this, ROB (2017) gives a more detailed explanation of the budgets of the municipalities. Engbersen et al. (2018) come up with possible factors that could explain the regional differences in the use of the three new laws. The factors include a cultural context (population mentality), an economic context (economic growth and stagnation), an institutional context (functioning of implementing organisations) and a physical context (shrinkage and urbanisation). Ooms et al. (2017) find that all these factors explain the three new laws significantly but that there are still opportunities for improvement of the model.

The research of Elissen and Ruwaard (2014) was commissioned by the Ministry of Health, Welfare and Sport. They have identified that experts agreed on thirteen population characteristics that are relevant to predict health care on individual level. It concerns five health characteristics: clinical status, functional status, complications of chronic illness, the use of care in a previous period and medication use. Furthermore, it concerns four personal characteristics: age, ethnicity, lifestyle and emotional concerns. In addition, it concerns four environmental characteristics: social-economical status, income, region and social network.

3 Data

This section explains the categorisation of the social care data of the Arrangementenmonitor, the data cleaning process, the demographic characteristics from the CBS data set and the construction of the target groups.

The research is commissioned by Thönissen Management en Advies B.V. (Thönissen, 2016).

They supplied social care data of 26 municipalities with 330 neighbourhoods on individual-, household-, neighbourhood- and municipality level. Furthermore, the demographic characteristics are extracted from the website of CBS (2020), the Dutch Central Office of Statistics. These data are available on neighbourhood- and municipality level.

The public sector uses different definitions for the social care facilities that they provide. This paper translates these social care facilities in English and provides a glossary in the Appendix to clarify these definitions for the Dutch municipalities.

3.1 Arrangementenmonitor

Municipalities use their own product names and product codes in their registration system. In addition, municipalities have four different domains which an individual service can belong to and each domain has its own way of data processing. This way, it is impossible to compare data between municipalities on product level. Therefore, Thönissen Management en Advies B.V. has developed the Arrangementenmonitor which classifies all possible product codes and product names to domains and sub-domains. The Arrangementenmonitor presents and analyses the social care data of all municipalities on individual- and household level. In addition, it can be used as a benchmark for the municipalities. As of 2016, the Arrangementenmonitor has been used by 26 municipalities for making and adjusting policy. Thönissen (2016) is convinced that social care domains are interrelated and therefore, the Arrangementenmonitor supports the integral approach. The integral approach considers all social care combined within a household, instead of addressing the problem for every domain separately.

Social care is divided into four domains: Social Support Act (SSA), Youth care, Education, and Participation. Table 1 shows that the Youth care domain can be categorised into four sub-domains. Furthermore, the table shows that the domain Participation can be categorised into the sub-domains Financial assistance and Special financial assistance. These are services for people that do not own enough money to make a living and for people that are faced with unforeseen expenses, respectively. Examples for Special financial assistance are costs for glasses and relocation. SSA facilities are social care facilities for people older than 18 years.

Examples of SSA facilities for elderly people are mobility facilities, help with living and housecleaning facilities.

In addition, Table 1 shows the division of social care into two categories: social mental care and social physical care. For the construction of target groups, it is important to distinguish between physical and mental social care because these are two different types of social care. Table 1 shows that SSA has five sub-domains of which Residential facilities and Transport SSA are defined by Thönissen Management en Advies B.V. as physical social care. People that need one of these two facilities are people that need help to physically function in society. The other social care facilities are labelled as mental social care since these facilities help people to function mentally in society.

3.2 Data cleaning

This subsection explains the data cleaning process that has been done to correct mistakes and to obtain a user-friendly data set. Although the Arrangementenmonitor is a satisfactory data set to use for research, a lot of data cleaning still needed to be done. The Arrangementenmonitor contains social care data of the years 2017, 2018 and 2019. It is presented per quarter such that $T = 12$. For these 12 quarters, the original data set contains 3,198,305 social care facilities with 249,992 clients, 186,757 households, 331 neighbourhoods and 26 municipalities. After data cleaning, the data set has been reduced to 2,817,503 social care facilities, 228,260 clients, 170,150 households, 330 neighbourhoods and 26 municipalities.

3.2.1 Incomplete reporting

The municipalities send their social care data to Thönissen Management en Advies B.V, which processes and uploads these data to the database in Tableau. Only some municipalities send the data of the domain Education and the sub-domains Student transport and Debt counseling to Thönissen Management en Advies B.V. In turn, the domain Education affects the two sub-domains Special education and Student transport. As a result, the reporting of the four sub-domains between municipalities is incomplete. Consequently, this gives biased results when constructing the target groups in the first analysis of this paper because the target groups are constructed by counting the number of distinct sub-domains in a household. Therefore, it has been decided to leave out these four sub-domains.

3.2.2 Age categories

Data is only available of the age categories of people that need social care. When two or more people within one household need social care, multiple age categories exist within that household. This poses the question of which age category to use. In order to tackle

Table 1: Categorisation of social care by the Arrangementenmonitor.

Domain	Sub-domain	Physical or Mental
Youth care	Guidance/daytime activities youth	Mental
	Youth mental health	Mental
	Living youth	Mental
	Remaining youth facilities	Mental
Participation	Financial assistance	Mental
	Special financial assistance	Mental
SSA	Daytime activities SSA	Mental
	Residential facilities	Physical
	Transport SSA	Physical
	Guidance SSA	Mental
	Living SSA	Mental

this question, the two age categories are merged together as one new age category. For example, a household with an eight year old and a fifty year old that both need social care are assigned the category of 'Young/Adults'. To construct the target groups in the end, these age categories are used. However, the problem arises because approximately 2,000 individuals are assigned two or more age categories within the same time period. This problem is for example caused when municipalities upload their data for 2017 in the year 2020. Sometimes, they forget to include that the social care facilities belong to the year 2017, instead of 2020. As a consequence, the year 2020 has been chosen and people incorrectly become 3 years older. Therefore, after consultation with Thönissen Management en Advies B.V, this research selects the youngest age category for people that obtained more than one age category in the same time period. From now on, this problem is automatically solved in the Arrangementenmonitor.

3.2.3 Neighbourhoods

The third analysis investigates which demographic characteristics explain social care on neighbourhood level. Around 5,000 households were classified to two or more neighbourhoods which is impossible because overlapping neighbourhoods do not exist. These mistakes are solved by individually looking at the zip code and reclassifying them to the correct neighbourhood. The procedure has now been changed in the Arrangementenmonitor, such that this problem will no longer occur.

3.2.4 Providers

Many providers for social care facilities exist and every department of the municipality inserts the names of the providers manually. As a result, a lot of mistakes are made. For example, the provider Zuyderland Thuiszorg is manually inserted in many ways, such as Zuyderland Zorg Thuis, Zuyderland Zorg, Zuiderland Thuiszorg, Zuijderland Thuis Zorg and Zuyderland Thuis. This way, the data set incorrectly shows that households with this provider need six different providers while, in reality, it is one and the same provider. At first, this research aimed to include the variable 'number of providers in a household' for the construction of the target groups. However, this resulted in too many target groups and therefore this variable has been excluded for this research. In order to solve the problem in the future, a procedure in SQL has been constructed that automatically corrects the wrong provider names.

3.2.5 Construction final data set

The Arrangementenmonitor is presented on individual-, household-, neighbourhood- and municipality level and the CBS data set is presented on neighbourhood- and municipality level. The demographic characteristics are not publicly available on household level and therefore the neighbourhood level is the preferred aggregated level. The extensive motivation for this is given in Section 2. The only common variable that the two data sets have is the neighbourhood variable. Based on this common variable, the data sets are merged together. The demographic characteristics are represented as percentages on neighbourhood level. Therefore, for every target group, a column is added that represents the percentage of households that belong to a certain target group. A hashing tool is used to pseudonymise the names of the neighbourhoods because this information is not allowed to be made public.

3.3 Demographic characteristics

The demographic characteristics from the CBS data set are shown in Table 2. This data set is categorised into seven categories: age category, income, consensual union, residence, household composition, nationality of immigrants and gender. The 29 demographic characteristics are represented as percentages on neighbourhood level, except the variable 'average income'. This variable considers only people that have an income and is represented in thousands of euros.

Table 2: The demographic characteristics per category represented as percentages on neighbourhood level.

Age category	**Income**	**Consensual union**
1. 0 - 3	10. Average income	16. Married
2. 4 - 11	11. 20% highest	17. Not married
3. 12 - 17	12. 40% lowest	18. Divorced
4. 18 - 22	13. High income	19. Widowed
5. 23 - 44	14. Low income	
6. 45 - 66	15. Prolonged low income	
7. 67 - 74		
8. 75 - 84		
9. 85+		

Residence	**Composition**	**Immigrants**	**Gender**
20. Average value	23. One person	26. Western	28. Men
21. Rental	24. With children	27. Non-Western	29. Women
22. Owner-occupied	25. No children		

The construction of the demographic characteristics classified under the category 'income' needs further explanation. CBS defines the variables 'low income' and 'high income' across households differently. The reason for this is that the net amount a household has to spend on an annual basis is adjusted for its size and composition. Regarding households with a 'low income', the boundary is 1,040 euros per month for a one person household and 1,380 euros for a household without children. The income of households with children is defined as low when it is less than 1,960 euros per month. One person households-, households without children- and households with children with a 'high income' have the boundaries 3,850, 5,270 and 7,480, respectively. If a household has a low income for at least four years, it is said to have a 'prolonged low income'. The composition of '20% highest incomes' and '40% lowest incomes' is performed in a different way. The Dutch households are sorted based on income. Thereafter, the amount of households that belong to the 20% highest incomes in the Netherlands and the households that belong to the 40% lowest incomes are counted per neighbourhood. Finally, these are divided by the total number of households per neighbourhood (CBS, 2020).

The nine age categories represent the percentage of a certain age group in a given neighbourhood. Furthermore, the percentages of 'married', 'not married', 'divorced' and 'widowed' are given in percentages per neighbourhood. Moreover, the residence characteristics are given per neighbourhood, being 'average residence value', 'rental properties' and 'owner-occupied properties'. In addition, the three types of composition of households are given in percentages, i.e. 'one person households', 'households with children', and 'households without children'. The percentage of immigrants per neighbourhood are categorised in 'Western' and 'non-Western'. Lastly, the percentage of 'men' and 'women' is given per neighbourhood.

3.4 Target groups

Motivated by related literature, as has been explained in Section 2, this research divides a population into distinct target groups since this contributes to a better targeted approach. The 27 target groups are presented in Table 3, as is the number of times the groups occur in the Arrangementenmonitor. Note that these are all possible distinct target groups in 12 quarters. Therefore, a household that changed target group is counted more than once. Supported with the background knowledge of Thönissen Management en Advies B.V, the target groups are constructed by using the following household characteristics: the distinct number of sub-domains (one, two, three, four or more), the classification of social care (physical, mental, both) and the age category (youth, youth/adults, adults, elderly people). Note that 'four or more' distinct sub-domains is abbreviated as '4+'. These characteristics are merged together as one string in the above order, separated by an underscore.

In case there is only one sub-domain, the name of this sub-domain is followed after the '1_' because this clarifies the target group and enhances a targeted approach. Table 1 gives an insight into the sub-domains that are classified as physical social care and mental social care. In case there is a combination of the age categories 'youth' and 'elderly' or 'adults' and 'elderly', Thönissen Management en Advies B.V chose to classify these to the category 'elderly' to reduce the number of target groups.

The three household characteristics give 4 x 3 x 4 = 48 combinations of target groups. Some target groups are too small and therefore merged together, as had been advised by Thönissen Management en Advies B.V. The difficulty of constructing the target groups is that, on the one hand, they should be constructed recognisably and, on the other hand, they should contain enough households to be able to conduct significant research. The expertise of Thönissen was needed to overcome this difficulty. An example of merged target groups are physical, mental, and physical/mental households that need three distinct sub-domains and 'elderly' as age category. These physical and mental groups are relatively small since a household with three sub-domains is usually a combination of physical- and mental social care. Thönissen is an expert in the field of social care and decides to merge these groups together because the target group '3_elderly' is clearly defined for municipalities. Furthermore, note that the target group 'Living SSA_Total' has merged all age categories into one age category.

Table 3: The target groups.

Target group	Number of households
1_Financial assistance_Adults	32,681
2_Mental_Adults	30,252
1_Youth mental health	27,829
1_Residential facilities_Elderly	19,608
1_Transport SSA_Elderly	18,772
2_Physical_Elderly	17,853
1_Special financial assistance_Adults	11,828
3_Elderly	10,005
1_Guidance SSA_Adults	9,899
2_Physical/Mental_Adults	8,748
2_Youth	8,568
1_Transport SSA_Adults	7,058
1_Residential facilities_Adults	6,854
3_Physical/Mental_Adults	6,289
2_Physical/Mental_Elderly	6,194
1_Guidance youth	5,897
3_Mental_Adults	5,828
4+_Adults	4,869
2_Youth/Adults	4,814
3_Youth/Adults	4,533
1_Remaining youth facilities	4,394
4+_Elderly	4,071
1_Living SSA_Total	3,344
4+_Youth/Adults	3,004
3+_Youth	2,235
1_Special financial assistance_Elderly	2,001
1_Guidance SSA_Elderly	1,314
Total	**170,704**

4 Comparing expert-constructed and data-driven target groups

This paper first examines whether the target groups, as defined in subsection 3.4, are statistically well-defined. Based on a clustering algorithm that clusters on the distinct number of sub-domains, this section constructs target groups in a data-driven way. Both sets of target groups are then compared using the Adjusted Rand Index to examine whether the target groups constructed by Thönissen are statistically justified. This analysis is performed in 'R' (R Core Team, 2020).

4.1 Methodology

4.1.1 K-means algorithm

The K-means algorithm is an iterative distance-based technique that partitions a given data set into a set of K clusters. The goal of this algorithm is to find groups in the data such that the variation inside the clusters is as small as possible and the variation between the clusters is as large as possible.

Anderberg (2014), Forgy (1965), Fukunaga (2013), MacQueen et al. (1967), Jain and Dubes (1988), Hartigan (1975), and Tou and Gonzalez (1974) report various methods to find clusters in a given data set but none of them are flawless. This paper focuses on the K-means algorithm of MacQueen et al. (1967) for three reasons. First, Bishop et al. (1995), Cheeseman et al. (1996), Ruiz et al. (2019) and Meilă and Heckerman (1998) state that it is one of the most used clustering algorithms that partitions a given data set into a set of K clusters. Second, Ruiz et al. (2019) and Meilă and Heckerman (1998) argue that the K-means algorithm is one of the simplest unsupervised learning algorithms that solves the well-known clustering problem. Third, the K-means algorithm can be used appropriately since two of the most important drawbacks have been overcome in this research. Among others, Verbeek (2002) and Fränti and Sieranoja (2018) claim that the biggest disadvantages of the K-means clustering algorithm are the requirement to pre-specify the number of clusters and the sensitivity to outliers. These two drawbacks are overcome because two acknowledged methods are used to investigate the number of clusters and the data set is relatively resistant to outliers, respectively. The reason for the data set being relatively resistant to outliers is because the sub-domains, on binary level, are clustered per household per quarter. This way, not the number of sub-domains but the number of distinct sub-domains are clustered. Certainly, there are some households that need more social care than other households and therefore need a higher distinct number of sub-domains. However, Thönissen Management en Advies B.V argues that households with seven or less distinct number of sub-domains are large-scale consumers of social care and should not be considered as outliers. Households with eight or more sub-domains (only 38 out of 111,639) are removed before using the clustering algorithm.

The K-means algorithm can be summarised in five steps (Wong and Hartigan 1979, MacQueen et al. 1967, Witten and Frank 2002, Han et al. 2011).

First, the number of K clusters needs to be defined. This paper discusses two acknowledged methods that investigate the optimal number of clusters. Among others, Marutho et al. (2018), Syakur et al. (2018) and Bholowalia and Kumar (2014) use the Elbow method to define the optimal number of clusters. The Elbow method calculates and plots the within-cluster sum of squares (WCSS) curve for a number of K clusters. The K clusters are graphed on the x-axis and the sum of squared errors are graphed on the y-axis. Thereafter, the accurate number of clusters is chosen visually at the location of a bend in the plot, assuming that at this point the sum of squared residuals is minimised. The Average Silhouette method is another method that is often applied (Rousseeuw 1987, Pollard and Van Der Laan 2002, Subbalakshmi et al. 2015) to investigate the optimal number of clusters. For a different number of K clusters, it measures the performance of each of the objects in the clusters. The silhouette width ranges from -1 to 1 where an average silhouette width of 1 corresponds to great performance: the objects are well matched to their own cluster and poorly matched to neighbouring clusters. Kaufman and Rousseeuw (1990) argue that the optimal number of clusters K is the one that maximises the average silhouette over a range of possible values for k. The different number of clusters are graphed on the x-axis and the average silhouette width is graphed on the y-axis.

Second, K number of objects are randomly chosen to be the centre of the K clusters.

Third, all remaining objects are partitioned into the K clusters based on the minimum squared-error criterion which measures the distance between an object and the cluster center. Formulated differently, all objects (x_j) are allocated to the closest cluster (S_i) with center μ_i, minimising the Euclidean distance to the cluster mean. This looks as follows:

$$min(WCSS(s)) = \sum_{i=1}^{k} \sum_{x_j \in S_i} || \; x_j - \mu_i \; ||^2 \tag{1}$$

where

- WCSS is the within-cluster sum of squares that should be minimised;
- x_j the object belonging to cluster S_i;
- μ_i the mean value of the objects assigned to cluster S_i.

Fourth, new mean cluster centers are calculated and updated for all K clusters.

Fifth, steps three and four are iterated until WCSS is minimised and the cluster centers do not change anymore. Note that the iteration process can stop earlier when the maximum number of iterations (10 by default) is reached.

4.1.2 Adjusted Rand Index

There is a large number of published studies (Steinley, 2004; Santos and Embrechts, 2009; Warrens, 2008; Yeung and Ruzzo, 2001) that use the Adjusted Rand Index (ARI) in cluster validation as a measure of agreement between two partitions. In this research, the two partitions are the specified target groups constructed with the help and expertise of Thönissen and the clusters that are constructed with the K-means algorithm.

It is assumed that, given a set S = $\{O_1, ..., O_n\}$ of n objects, U = $\{u_1, ..., u_{27}\}$ represents the set of the 27 target groups and V = $\{v_1, ..., v_K\}$ represents the set of K clusters. Furthermore, it is assumed that both U and V represent two different partitions of the objects in S, being a partition of S into 27 subsets and a partition of S into K subsets. Rand (1971) formulates the Rand Index simply as

$$\frac{a+d}{a+b+c+d} \tag{2}$$

where

- a is the number of pairs of objects in S that are in the same subset in U and in the same cluster in V;
- b is the number of pairs of objects in S that are in the same subset in U but in different clusters V;
- c is the number of pairs of objects in S that are in the same cluster in V but in different subsets in U;
- d is the number of pairs of objects in S in different subsets and different clusters in both partitions.

Steinley (2004) assumes that, for the ARI, the U and V partitions are picked randomly such that the number of objects in the subsets and clusters are fixed. Vinh et al. (2010) state that the ARI is the chance corrected version of the Rand Index by using random clustering.

The boundaries of the ARI are -1 and 1, where the number 1 indicates that the two partitions are exactly the same. Scrucca et al. (2016) report in their 'R' package the following definitions with corresponding boundaries: ARI ≥ 0.90 excellent recovery; $0.80 \leq$ ARI < 0.90 good recovery; $0.65 \leq$ ARI < 0.80 moderate recovery; ARI < 0.65 poor recovery.

4.2 Empirical results

This section discusses the results of the K-means clustering algorithm. Furthermore, by using the Adjusted Rand Index (ARI), it examines whether the constructed target groups of subsection 3.4 are statistically well-defined. This is done by comparing these to target groups that have been constructed by the clustering algorithm. This way, the categorisation of the variables 'number of distinct sub-domains', 'type of social care', and 'age category of the household' are compared with the combination of the different types of distinct sub-domains.

The clustering algorithm is performed on the fourth quarter of 2018 for two reasons. First, this time period considers the highest number of households. Second, this research wants to allocate the same weight factor to all households and target groups, thus focusing on one quarter only. A binary vector is made for every sub-domain, consisting of a 0 if the household did not need the given sub-domain and a 1 if the household did need the given sub-domain.

Following the five steps mentioned in subsection 4.1.1, the optimal number of clusters has to be defined first. Figure 1 and Figure 2 show the results of the Elbow method and the Silhouette method.

The location of a bend, also known as the elbow, is visible at k = 4 and k = 7 in Figure 1. The average silhouette width is highest at k = 24, with a silhouette width of 0.67. However, the cluster k = 19 almost report an equally high silhouette width and therefore it does not choose k = 24 distinctly. This is a reason to choose the Elbow method. On the other hand, a reason to favour the Silhouette method over the Elbow method is that the last one is more open to interpretation.

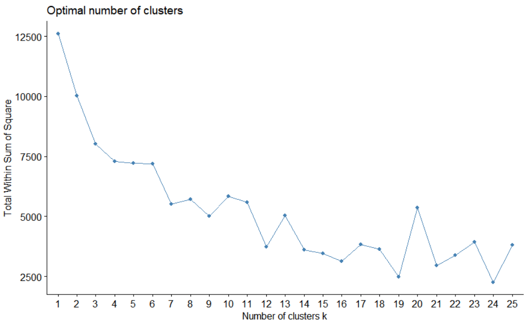


Figure 1: The Elbow method.

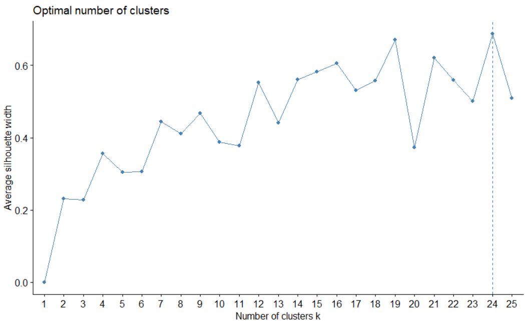


Figure 2: The Silhouette method.

Following the Elbow method, the K-means algorithm is executed for k = 4 and k = 7 and every household is categorised to clusters 1 until 4 and 1 until 7, respectively. The 27 target groups are compared to the 4 and 7 clusters. The ARI scores report 0.20166 and 0.42771 which both correspond to poor recovery.

Using the Silhouette method, the K-means algorithm is executed for k = 19 and k = 24 and every household is categorised to clusters 1 until 19 and 1 until 24, respectively. The 27 target groups are compared to the 19 and 24 clusters. The ARI scores of 0.62281 and 0.67331 correspond to poor and moderate recovery, respectively.

The results are summarised in Table 4. To conclude, the Silhouette method clusters the distinct sub-domains in 24 clusters. The ARI score shows that the constructed target groups in subsection 3.4 are statistically well-defined when taking into account the combination of the different types of distinct sub-domains. Therefore, these target groups are used for the remainder of this research.

Table 4: Results of the Elbow- and Silhouette method.

Method	Number of clusters	ARI	Confidence interval
Elbow method	4	0.20166	[0.20128, 0.20204]
Elbow method	7	0.42771	[0.42749, 0.42793]
Silhouette method	19	0.62281	[0.62250, 0.62313]
Silhouette method	24	0.67331	[0.67301, 0.67361]

5 Dynamics of the target groups

This section gives more insight into the dynamics of the target groups. That is, the in-, out- and through-flow for and between the target groups. This is the first study to undertake an analysis of the dynamics of social care, thus contributing to a deeper understanding of the patterns of households that need social care. This analysis is implemented using the statistical software tool Tableau (2020).

5.1 Methodology: Tableau procedure

This subsection briefly explains the code that has been constructed in Tableau for the analysis of the in-, out- and through-flow of the target groups.

First, a parameter is constructed that enables municipalities to choose the year and quarter of interest. Second, two calculated fields, named 'T/F date of interest' and 'T/F one quarter before date of interest', are constructed that either return a 'true' or a 'false'. The 'true' values are the ones of interest and correspond to the selected time period of the parameter. Third, two calculated fields called 'target group date of interest' and 'target group one quarter before date of interest' are created. They return the corresponding target group in case the 'T/F date of interest' and 'T/F one quarter before date of interest' return a 'true', respectively.

After the first three constructions, another calculated field is created that considers four different cases. A challenging 'if-statement' is constructed which ensures that the cases are only considered when the target groups belong to the period of interest. Formulated differently, the cases are not considered when the fields 'target group date of interest' and 'target group one quarter before date of interest' are both empty. The procedure of the four different cases is the following. First, if the field 'target group date of interest' is empty, it means that there was an out-flow of the target group. Second, if the field 'target group one quarter before date of interest' is empty, it means that there was an in-flow of the target group. Third, if the two fields are equal to each other, it means that there was no in-, out- or through-flow and therefore the target group remains the same. This is defined as 'no mutation'. Fourth, if the 'if-statement' was true and the above three cases were 'false', there must have been a through-flow since no other options are left. The through-flow is defined as the sum of in-through-flow and out-through-flow, as will be explained in the next subsection.

5.2 Empirical results

This subsection discusses the most striking results of the in-, out- and through-flow. The complete dashboards for the in-, out- and through-flow are constructed in Tableau and available online for all municipalities of the Arrangementenmonitor. This way, all results can be accessed in Tableau. The interface can be used dynamically and the data of interest can be easily filtered to the needs of the user. For example, municipalities can investigate the in-, out- and through-flow for every neighbourhood, target group and time period.

Table 5 reports the in-, out- and through-flow from the third to the fourth quarter of 2018 (2018 Q4) of every target group for all participating municipalities of the Arrangementenmonitor. To put numbers in context, the in-, out- and through-flow can be divided per row by 'no mutation'. The average fractions of the in-, out- and through-flow compared to no mutation are 6.1%, 6.1% and 14.6%, respectively.

The following target groups show a relatively large in-flow (more than 12,5%) in 2018 Q4: 1_Special financial assistance_Adults, 1_Special financial assistance_Elderly, 1_Youth mental health and 1_Guidance SSA_Elderly. The target group 1_Special financial assistance_Adults reports the highest in-flow fraction with 26.2%.

The following target groups show a relatively large out-flow (more than 12,5%) in 2018 Q4: 1_Guidance youth, 1_Special financial assistance_Elderly, 1_Residential facilities_Adults, 1_Youth mental health and 1_Guidance SSA_Elderly. The target group 1_Special financial assistance_Adults reports the highest out-flow fraction with 22.8%.

The target group 1_Financial assistance_Adults and all 2+ target groups except 2_Physical_Elderly show a relatively large through-flow (more than 25%) in 2018 Q4. Note that the target groups 2_Youth (103.4%) and 3_Youth/Adults (130.6%) even have more through-flow than 'no mutation'.

In summary, the first two results make intuitively sense because the highest in- and out-flow most often occur in households that need one sub-domain. The third result provides important new insights in terms of the through-flow. Relatively more through-flow occurs in

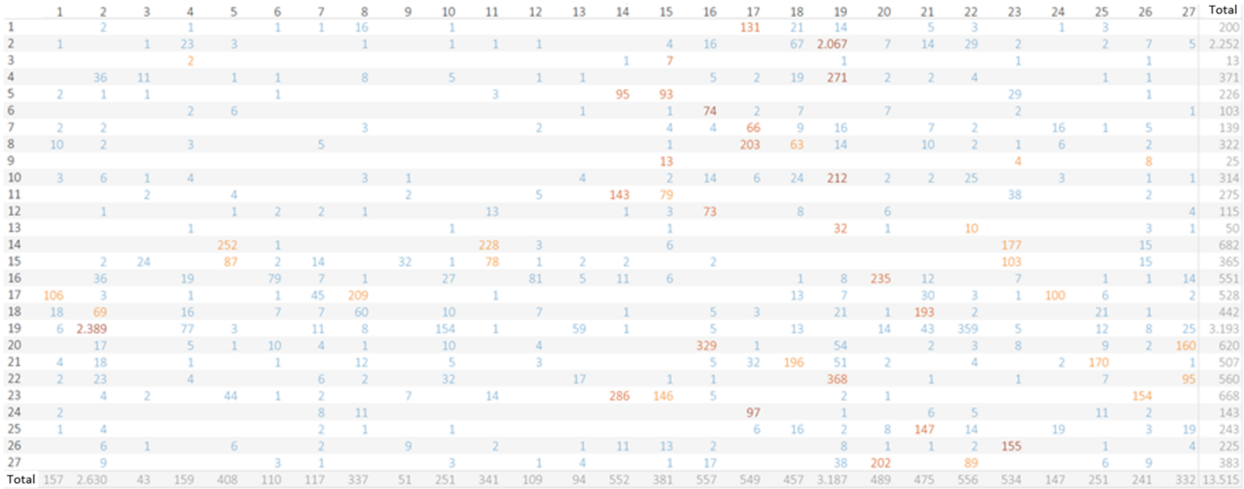
Table 5: The in-, out- and through-flow in the fourth quarter of 2018.

Target group	In-flow	Out-flow	Through-flow	No mutation
1. 1_Guidance youth	179	249	357	1,465
2. 1_Financial assistance_Adults	634	1,211	4,882	11,678
3. 1_Special financial assistance_Elderly	110	109	56	699
4. 1_Special financial assistance_Adults	820	714	530	3,133
5. 1_Residential facilities_Elderly	469	300	634	9,061
6. 1_Residential facilities_Adults	167	82	213	1,681
7. 1_Remaining youth facilities	121	112	256	1,519
8. 1_Youth mental health	1,389	1,406	659	9,247
9. 1_Guidance SSA_Elderly	56	45	76	359
10. 1_Guidance SSA_Adults	323	302	565	3,409
11. 1_Transport SSA_Elderly	550	258	616	9,917
12. 1_Transport SSA_Adults	139	80	224	3,802
13. 1_Living SSA_Total	52	71	144	898
14. 2_Physical_Elderly	107	103	1,234	8,431
15. 2_Physical/Mental_Elderly	78	70	746	1,875
16. 2_Physical/Mental_Adults	65	42	1,108	3,367
17. 2_Youth	61	62	1,077	2,388
18. 2_Youth/Adults	13	15	899	869
19. 2_Mental_Adults	249	309	6,380	6,631
20. 3_Physical/Mental_Adults	13	14	1,109	1,983
21. 3_Youth/Adults	9	5	982	752
22. 3_Mental_Adults	7	18	1,116	1,316
23. 3_Elderly	17	54	1,202	3,516
24. 3+_Youth	9	8	290	658
25. 4+_Youth/Adults	3	1	494	940
26. 4+_Elderly	1	9	466	1,213
27. 4+_Adults	3	11	715	1,673
Total	**5,644**	**5,660**	**13,515**	**92,480**

households that need two or more sub-domains. Two reasons for this could be these households having a higher chance of a) needing an extra sub-domain or b) not needing one of their sub-domains anymore.

This paper defines the in-through-flow as the in-flow from another target group and the out-through-flow as the out-flow to another target group. The through-flow is defined as the sum of the in-through-flow and the out-through-flow. It is interesting to examine the through-flow in more detail because it provides a deeper understanding of the mutations of social care within households.

Figure 3 presents the through-flow between target groups from 2018 Q3 to 2018 Q4. It shows the visualisation built in Tableau that is available for all participating municipalities of the Arrangementenmonitor. The first column represents the target groups of 2018 Q3 and the first row represents the target groups of 2018 Q4. The last column, which sums the columns per row, represents the in-through-flow. The last row, which sums the rows per column, represents the out-through-flow. The numbers in the first row and first column correspond to the numbers of the target groups of Table 5. Numbers in blue and red represent a low and a high share of the in-through-flow, respectively.



	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	Total
1		2		1		1	1	16		1							131	21	14		5	3		1	3			200
2	1		1	23	3			1		1	1	1			4	16		67	2.067	7	14	29	2		2	7	5	2.252
3				2										1	7				1				1			1		13
4		36	11		1	1		8		5		1	1			5	2	19	271	2	2	4			1	1		371
5	2	1	1			1					3			95	93								29			1		226
6				2	6								1		1	74	2	7		7			2				1	103
7	2	2						3				2			4	4	66	9	16		7	2		16	1	5		139
8	10	2		3			5								1		203	63	14		10	2	1	6		2		322
9															13								4			8		25
10	3	6	1	4				3	1				4		2	14	6	24	212	2	2	25		3		1	1	314
11			2		4				2			5		143	79								38			2		275
12		1			1	2	2	1			13			1	3	73		8		6							4	115
13				1						1					1				32	1		10				3	1	50
14					252	1					228	3			6								177			15		682
15		2	24		87	2	14		32	1	78	1	2	2		2							103			15		365
16		36		19		79	7	1		27		81	5	11	6			1	8	235	12		7		1	1	14	551
17	106	3		1		1	45	209			1							13	7		30	3	1	100	6		2	528
18	18	69		16		7	7	60		10		7		1		5	3		21	1	193	2			21	1		442
19	6	2.389		77	3		11	8		154	1		59	1		5		13		14	43	359	5		12	8	25	3.193
20		17		5	1	10	4	1		10		4				329	1		54		2	3	8		9	2	160	620
21	4	18		1		1		12		5		3				5	32	196	51	2		4		2	170		1	507
22	2	23		4			6	2		32			17		1	1			368		1		1		7		95	560
23		4	2		44	1	2		7		14			286	146	5			2	1						154		668
24	2						8	11									97		1		6	5			11	2		143
25	1	4					2	1		1							6	16	2	8	147	14		19		3	19	243
26		6	1		6		2		9		2		1	11	13	2			8	1	1	2	155		1		4	225
27		9				3	1			3		1	4		1	17			38	202		89			6	9		383
Total	157	2.630	43	159	408	110	117	337	51	251	341	109	94	552	381	557	549	457	3.187	489	475	556	534	147	251	241	332	13.515

Figure 3: The through-flow between target groups.
This dashboard is built in Tableau and shows the through-flow between target groups from 2018 Q3 to 2018 Q4. The numbers in blue and red represent a low and a high share of the in-through-flow, respectively.

Figure 3 shows some noteworthy results, especially for 2_Mental_Adults (group 19). Of the 3,193 in-through-flow of group 19, 2,389 come from the target group 1_Financial assistance_Adults (group 2). Therefore, considering the age category 'adults' in 2018 Q4, 75% of all new in-through-flow of group 19 comes from the target group that only needed the sub-domain financial assistance in 2018 Q3. As a result, a household needing financial assistance in 2018 Q3, relatively often needed another sub-domain in 2018 Q4. The same pattern is visible the other way around; of the 2,252 in-through-flow of group 2, 2,067 come from group 19 (92%). Furthermore, 73% of the in-through flow of 1_Special financial assistance_Adults (group 4) comes from group 19. In addition, 68% of the in-through flow of 1_Guidance SSA_Adults (group 10) comes from group 19. In conclusion, there is a relatively high through-flow from target groups that have two distinct sub-domains in 2018 Q3 to the three target groups that have one distinct sub-domain in 2018 Q4. This result suffices for the municipalities since their policy aims at decreasing the number of distinct sub-domains within a household (Senate, 2014).

Three more striking results are present which show more than 67% in-through-flow. First, 74 out of 103 (72%) of the in-through-flow of 1_Residential facilities_Adults (group 6) come from 2_Physical/Mental_Adults (group 16). Again, this is a satisfactory result for the municipalities because of the decrease in distinct sub-domains. Second, 97 out of 143 (68%) of the in-through-flow of 3+_Youth (group 24) come from 2_Youth (group 17). Third, 155 out of 225 (69%) of the in-through-flow of 4+_Elderly (group 26) come from 3_Elderly (group 23). The second and third result indicate a relatively high increase in the number of distinct sub-domains for a) households with the age category 'youth' needing two distinct sub-domains and b) households with the age category 'elderly' needing three distinct sub-domains. The current research explores, for the first time in literature, that more attention is needed for these type of households.

Another dynamic dashboard that was built in Tableau is presented in Figure 4. In this figure, the numbers change according to the interest of the municipalities. They have the opportunity to choose every target group, neighbourhood and time period they are interested in. For example, Figure 4 visualises the through-flow of target group 19 for all municipalities in 2018 Q4. The y-axis graphs the number of households and the x-axis graphs two time periods. The top row corresponds to 2018 Q4 and the bottom row corresponds to 2018 Q3. For example, the highest light red bar corresponds to the through-flow of 2,389 households from group 2 in 2018 Q3 to group 19 in 2018 Q4. From this figure can be noted that the absolute number of in-through-flow and out-through-flow between groups 2 and 19 is relatively high. Adults needing financial assistance in 2018 Q3 relatively often need an extra

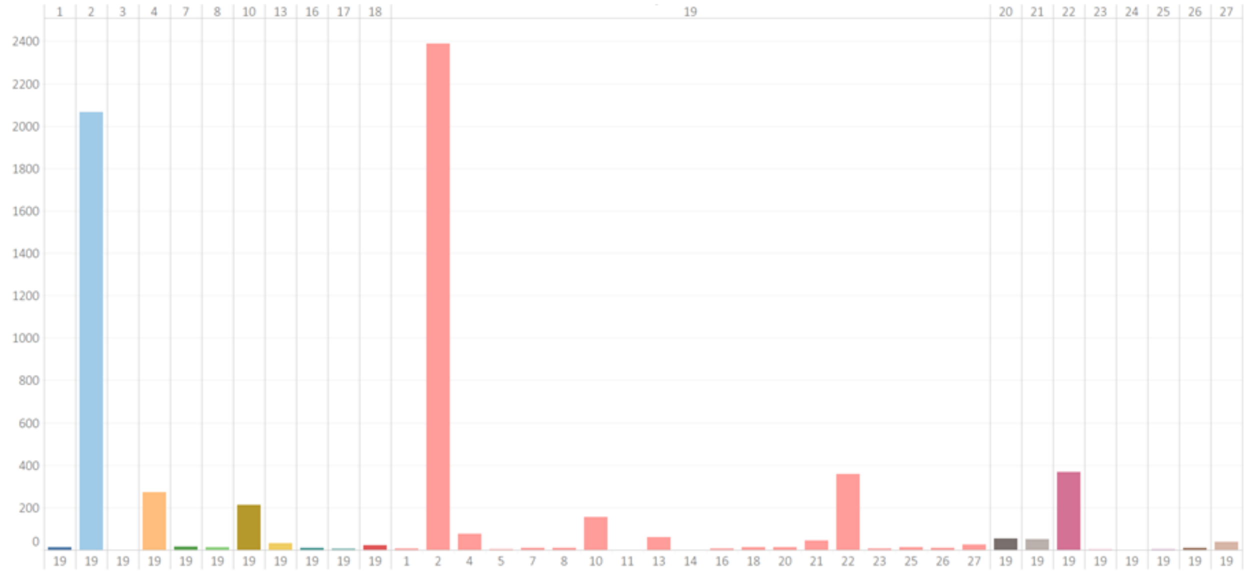


Figure 4: A closer look at the through-flow of target group 19.
This dashboard, built in Tableau, visualises the through-flow of households with age category 'adults' that need two distinct sub-domains of the type 'physical'. The x-axis graphs the two time periods 2018 Q4 (top row) and 2018 Q3 (bottom row). The y-axis graphs the number of households.

mental sub-domain in 2018 Q4 and vice versa. Municipalities have to consider these households and investigate possible explanations for this.

6 Demographic characteristics of neighbourhoods

This section examines which demographic characteristics significantly explain each of the 27 expert-constructed target groups. The first subsection explains the methodology and the second subsection discusses the results. The analysis is carried out using the statistical software tool Stata (2020).

6.1 Methodology

This subsection describes the methods and approaches that were used in the third analysis. It clarifies the variable- and model selection and it discusses the final model.

6.1.1 Variable selection

This subsection explains the selection of the 29 demographic characteristics from the CBS data set, shown in Table 2. Within this data set, some variables are highly correlated, causing multicollinearity in the model. Multicollinearity reduces the reliability of the predicted coefficients because of variables partially overlapping. Moreover, when including all variables of a certain category, for example men and women, perfect multicollinearity occurs. The reason for this is the value of men being able to predict the value of women. Multicollinearity is a problem because it undermines the statistical significance of an independent variable and ordinary least squares cannot be computed (Allen, 1997). Therefore, one must be careful for this phenomenon.

The Pearson correlation coefficients (Benesty et al., 2009) were computed to test for high correlations. The results, in percentages, are shown in Table 6. The numbers in bold in the first row and first column correspond to the numbers of the variables of Table 2. If the variables cause multicollinearity, the background is presented in light grey. The variables selected for the final model have a green background colour. It can be noted from the table below that, for example, the variables of the 'residence' category and the variable 'non-Western immigrants' are highly correlated with the variables of the 'income' category, thus causing multicollinearity. In addition, the variables of the 'income' category are highly correlated with each other, except for variables 13 and 14. In conclusion, several variables report correlations higher than 80% which results in multicollinearity. Therefore, the following independent variables (in percentages per neighbourhood) are excluded from the model: average income, 20% highest income, 40% lowest income, prolonged low income, not married, average residence value, rental houses, owner-occuppied houses, households without children, non-Western immigrants and women.

The age categories retrieved from the CBS data set do not provide new insights for this research, since the dependent variable is constructed using the age categories of households. As a result, they are not included in Table 6. However, it could be interesting to investigate whether, for example, the 85+ age category or the 67-74 age category explains more of a given target group with age category 'elderly'. Therefore, only the 'extreme' age categories are selected in the final model. In addition, the age category of students (18-22) is included in the model, since Thönissen Management en Advies B.V expects this variable to be significantly important. The reason for this is the expected low need of social care in neighbourhoods with a high presence of students because students are assumed to be the healthier group of the population.

The following independent variables (in percentages per neighbourhood) are selected in the final model: the age categories 0-3, 18-22 and 85+, high income, low income, married, divorced, widowed, one person households, households with children, Western immigrants and men. In the end, this resulted in a model without multicollinearity influencing the correctness of the results since these variables have correlations lower than 80%.

Table 6: The correlations between the variables.

	10	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**	**21**	**22**	**23**	**24**	**25**	**26**	**27**	**28**
10	100																		
11	85	100																	
12	-27	-80	100																
13	35	96	-83	100															
14	-30	-78	88	-67	100														
15	-33	-81	89	-84	93	100													
16	5	53	-80	55	-72	-73	100												
17	1	-34	63	-36	56	58	-95	100											
18	-27	-60	54	-62	63	64	-22	-7	100										
19	-5	-9	-7	-6	-20	-19	41	-63	27	100									
20	53	73	-50	84	-58	-56	15	0	-56	-6	100								
21	-28	-83	91	-84	89	89	-76	56	64	-5	-54	100							
22	28	83	-92	84	-88	-87	76	-57	-63	3	53	-99	100						
23	-6	-66	88	-68	70	71	-79	78	32	-12	-20	80	-81	100					
24	6	63	-80	63	-56	-57	74	-58	-31	-9	17	-72	73	-81	100				
25	5	49	-75	50	-72	-72	91	-87	-24	44	21	-70	70	-86	54	100			
26	-5	-46	51	-49	44	46	-35	24	33	2	-37	53	-53	51	-50	-37	100		
27	-33	-54	53	-55	82	86	-47	37	48	-23	-45	64	-62	39	-21	-58	28	100	
28	-41	12	-28	11	-16	-14	27	-17	-14	-21	-13	-23	23	-36	34	29	-11	10	100

The correlations are shown in percentages. The numbers in bold in the first row and first column correspond to the numbers of the variables of Table 2. If the variables cause multicollinearity, the background is presented in light grey. The variables selected for the final model have a green background colour.

6.1.2 Model selection

Different methods have been proposed to analyse panel data. Among others, Baltagi (2008) and de Crombrugghe (2020) propose to use the pooled ordinary least squares (POLS) model, the pooled ideal- or pooled feasible generalised least squares (PIGLS/PFGLS) model or the pooled instrumental variable (PIV) model for panel data without heterogeneity. In addition, they propose the random effects model or the fixed effects model for panel data with heterogeneity. In statistics, heterogeneity is known as 'individual effects'. This research assumes that the neighbourhoods of the Arrangementenmonitor are unique, contain individual effects and should not be treated equally. Therefore, because of the presence of heterogeneity, a random- or a fixed effects model is favoured. Throughout the remainder of the paper, when discussing the fixed effects model, it refers to the within estimator version of the fixed effects model.

The choice of the random- or fixed effects model depends on whether the heterogeneity η_i is correlated with the independent variables. In case η_i is viewed as a component of the error term ϵ_{it}, it is called a 'random effect'. In case η_i is estimated as an unknown individual-specific parameter, it is called a 'fixed effect'. The random effects assumption of Mundlak (1978) is an orthogonality assumption that states that the unobserved heterogeneity η_i is uncorrelated with the independent variables x_{it}:

$$Cov(\eta_i, x_{it}) = 0 \; \forall \; t \tag{3}$$

The Hausman test examines whether the heterogeneity is correlated with the independent variables. Wooldridge (2010) and Cameron and Trivedi (2005) state that the comparison of the random- and fixed effects models only concerns the effects of time-varying independent variables. Therefore, the standard model

$$y_{it} = x'_{i,t} \; \beta + \eta_i + \epsilon_{it} \tag{4}$$

can be rewritten as

$$y_{it} = \tilde{x}'_{it} \; \delta + w'_i \; \gamma + \eta_i + \epsilon_{it} \tag{5}$$

where

- x'_{it} is a subvector of $x_{i,t}$ containing only those independent variables that vary over time;
- w'_i is a subvector of $x_{i,t}$ containing the (remaining) independent variables that are constant over time, including the intercept;
- η_i is the heterogeneity;
- ϵ_{it} is the error term.

The aim of the Hausman test (Hausman, 1978) is to compare the effects of the time-varying independent variables δ of the fixed effects (δ_{fe}) and random effects (δ_{re}) model.

Hausman's test statistic is a quadratic form of the difference between δ_{fe} and δ_{re} and is calculated using the following formula:

$$H = \frac{[\hat{\delta}_{fe} - \hat{\delta}_{re}]' \, [\hat{\delta}_{fe} - \hat{\delta}_{re}]}{A_{var}(\hat{\delta}_{fe}) - A_{var}(\hat{\delta}_{re})} \overset{a}{\sim} \tilde{\chi}^2_k \tag{6}$$

where A_{var} is the asymptotic variance of the estimator. The asymptotic variance tries to approach the true value of the parameter when the limit is taken. Greene (2011) defines the asymptotic variance in the following way:

$$A_{var}(\hat{\theta_n}) = \frac{1}{n} \lim_{n\to\infty} \mathbb{E}[\{\sqrt{n}(\hat{\theta_n} - \lim_{n\to\infty} \mathbb{E}[\theta_n])\}^2] \tag{7}$$

Cameron and Trivedi (2010) explain that, under the null hypothesis of the Hausman test, $\hat{\delta}_{fe}$ and $\hat{\delta}_{re}$ are both consistent but $\hat{\delta}_{re}$ is the only efficient parameter.

6.1.3 The model

The final model is represented in the following equation:

$$\begin{aligned} TG_{i,t} = \beta_0 &+ \beta_1 * A11_{i,t} + \beta_2 * A17_{i,t} + \beta_3 * A74_{i,t} + \beta_4 * A84_{i,t} + \beta_5 * HI_{i,t} \\ &+ \beta_6 * LI_{i,t} + \beta_7 * MAR_{i,t} + \beta_8 * DIV_{i,t} + \beta_9 * WID_{i,t} \\ &+ \beta_{10} * OPH_{i,t} + \beta_{11} * HWC_{i,t} + \beta_{12} * WI_{i,t} + \beta_{13} * M_{i,t} + \eta_i + \mu_t + \epsilon_{i,t} \end{aligned} \tag{8}$$

where, in neighbourhood i = 1, .., 330 and at time period t = 1,..,12:

- $TG_{i,t}$ is the % of target group;
- β_0 is the constant;
- $A11_{i,t}$ is the % of people with the age between 4 and 11;
- $A17_{i,t}$ is the % of people with the age between 12 and 17;
- $A74_{i,t}$ is the % of people with the age between 67 and 74;
- $A84_{i,t}$ is the % of people with the age between 75 and 84;
- $HI_{i,t}$ is the % of households with a high income;
- $LI_{i,t}$ is the % of households with a low income;
- $MAR_{i,t}$ is the % of people that is married;
- $DIV_{i,t}$ is the % of people that is divorced;
- $WID_{i,t}$ is the % of people that is widowed;
- $OPH_{i,t}$ is the % of one person households;

- $HWC_{i,t}$ is the % of households with children;
- $WI_{i,t}$ is the % of Western immigrants;
- $M_{i,t}$ is the % of men;
- η_i is the individual effect;
- μ_t is the time fixed effect;
- $\epsilon_{i,t}$ is the error term.

Note that there are K = 27 target groups of which it has been examined whether they are explained by the demographic characteristics. Accordingly, there are 27 dependent variables and research is conducted on 27 models.

6.2 Empirical results

This section explains the choice for the fixed effects model and it discusses the most important findings of the analysis of the demographic characteristics.

The 27 random effects models and 27 fixed effects models are compared with each other. The Hausman test shows that, in all 27 cases, the unobserved heterogeneity (η_i) is correlated with the independent variables x_{it} by reporting a p-value always below 0.0001. According to the Bonferroni correction (Weisstein, 2004), to counteract the problem of multiple comparisons, the level of significance for hypothesis testing must be divided by the number of hypotheses, which is 27 in this research. At significance level $\alpha = 0.01$, this gives a Bonferroni corrected value of $\alpha* = 0.00037$. Consequently, the null hypothesis is rejected and the fixed effects model is chosen for examining whether the demographic characteristics explain the 27 target groups.

The results obtained from the third analysis are presented in Table 7. The independent variables are shown in the first row of the table. As shown, the variables either positively (+) or negatively (-) influence the target groups. From the asterisks can be noted whether the positive or negative influence is significant. The numbers in the first column of the table correspond to the numbers of the target groups in Table 5.

The goodness-of-fit measure ranges from 1.3% to 30.7% across the different target groups. It is lowest for households with adults needing three mental sub-domains (group 22) and it is highest for households with adults needing the sub-domain 'residential facilities' (group 6). This means the demographic characteristics explain target group 22 relatively poorly whereas they explain target group 6 relatively well. A reason for this difference is target group 22 considering a larger amount of sub-domains, thus having a higher variability inside the target group. The R^2 of the model of target group 6 is the only one scoring higher than 30%. The low R^2 values of the 27 models are not surprising since several independent variables explaining the need for social care are not taken into account.

From Table 7 can be noted that the variables across the different target groups are significant at least 4 out of 27 times at a 1% significance level. In general, the variables share their positive signs across the groups which means that social care consumption increases as the variables increase, ceteris paribus. The remainder of this section discusses the results of the most important variables.

Table 7: An overview of the results of the third analysis: Which demographic characteristics explain the target groups?

TG	A11	A17	A74	A84	HI	LI	MAR	DIV	WID	OPH	HWC	WI	M	R^2
1.	+	+	+*	+	+*	+*	+	-	+**	-**	-*	+	-**	0.072
2.	-**	-*	-*	-*	-*	+	+*	-*	+**	-	+	-*	-	0.154
3.	+**	-	-	-	-	+**	+*	+	+	+	+	+	+	0.019
4.	+*	+**	+*	+	-*	+*	-*	+	-**	-	-	-	+	0.032
5.	-*	+**	+*	+*	+*	+*	+*	+*	-**	+	-	-	+*	0.190
6.	+	+	+*	+*	+*	+*	-**	+*	-*	+*	-	-	-	0.355
7.	+**	-*	+*	+*	+*	+*	-*	+	+	+	+	+	-	0.057
8.	-**	+*	+*	+	+*	+*	-	+	-	+*	+*	-	+*	0.184
9.	-*	+	+**	+	+*	+	+*	+	+*	+	+	+**	+*	0.029
10.	-	+	+*	+*	-*	+*	-*	+*	+*	-	-	+*	+*	0.255
11.	+*	+*	+*	+*	+*	+*	-*	-**	+*	+	-*	+	-	0.257
12.	+*	+*	+*	+*	-*	+*	-*	+	-	+	-**	-	+	0.245
13.	+	+	-*	-	+	-	+	+	+	+*	+*	-**	+	0.019
14.	-	+	+	+*	+*	+*	-*	+*	-	+**	+*	+	-	0.201
15.	-	-	-	+**	+*	+**	-	+	+	-	-	+	-	0.017
16.	+	+	+*	+*	+*	+*	-*	+*	-	+	+	+**	-	0.157
17.	+	+*	+*	+*	+*	+	-*	+*	-	-**	-*	-*	-*	0.177
18.	+	+*	+*	+	+*	+	-*	-*	+	+	-	+	-	0.032
19.	-	-	-*	-	-	+	+	-	+**	-**	+	-	-	0.021
20.	+**	+*	+*	+*	+*	-	-**	+*	+	-	-*	+	-	0.062
21.	+	+	+*	+	+*	+	-*	+	+**	-	-	-	+	0.034
22.	+	-	+*	+	+	-	+	+*	-	-	-	+	+	0.009
23.	+	+*	+*	+*	+*	+*	+**	+*	+*	+	*-	+*	-	0.101
24.	+	+	+*	+*	+*	+*	-*	+*	+	-*	-*	-	+	0.111
25.	+	+*	+*	+	+*	+*	-	+	+	+	-**	+*	+*	0.073
26.	-	+	-	+	+**	-	+	-*	-	-*	-	-*	-	0.019
27.	+	+*	+*	+*	+*	+	-	+*	+	+	+	+*	+*	0.063
All.	+	+*	+*	+*	+*	+*	-*	+*	-	+	-	-**	+	0.338

** p-value < 0.05; * p-value < 0.01
The numbers in the first column correspond to the numbers of the target groups in Table 5.
The demographic characteristics are abbreviated. One after another, they are: age categories 4-11, 12-17, 67-74 and 75-84, high income, low income, married, divorced, widowed, one person household, household with children, Western immigrants, men.

Surprisingly, significant differences between target groups were found for the variable married. In some cases, the percentage of married people positively influences the target groups and in other cases it negatively influences the target groups. Especially the number of target groups with the age category ‘youth’ increases as the percentage of married households in a neighbourhoods decreases. A possible explanation is that social youth care is needed more often when the parents of the child needing social care are not married. A further study is suggested, with a stronger focus on this variable, in order to investigate what effect being married has on social care consumption.

It is interesting to note that in most cases of this study, target groups need more social care when the percentage of divorced people in a neighbourhood is higher. The variable ‘divorced’ confirms the hypothesis that social youth care is needed more often when the parents of the child needing social care are not married: the higher the percentage of divorced people in a neighbourhood, the higher the need for social youth care.

Another important finding is that a household with a low income almost always significantly explains the target groups. The plus-signs in Table 7 in the columns of ‘LI’ correspond to more social care in every target group when a household has a low income.

Strong evidence was found for the fact that households with a high income need less financial help and households with a low income need more financial help. This is an expected result since only households with a low income are the ones that are qualified for financial help. The most surprising aspect of the analysis is the positive influence of households with a high income for almost all remaining target groups. In some cases, this influence is even significant. A possible explanation for ‘Youth mental health’ being significant with the variable ‘high income’ could be rich people being less hesitant to go to a doctor for mental health issues. Consequently, they are more often referred to a specialist for youth mental health services. Further research should be undertaken to investigate whether this hypothesis is supported.

Considering the positive influence of households with a high income for almost all remaining target groups, in particular, a positive correlation was found between target groups with the age category ‘elderly’ and the variable ‘high income’. Interestingly, this result means that especially elderly people with a high income significantly need more social care. According to CBS (2009), people with a low income live about five years less than people with a high income. Therefore, there are relatively many people with a higher income in the highest age category.

Target groups with three or more social care facilities - divided over 330 neighborhoods - do not appear in significant amounts to draw reliable conclusions about whether the demographic characteristics of a neighborhood explain these target groups. Further research is required to gain more insight into which demographic characteristics explain these target groups.

This research performs sensitivity analysis such that it is examined whether the most important results still hold when correlated variables are interchangeably used in different models. Subsection 6.1.1 has explained the omission of some variables in the model because they were causing multicollinearity. However, it is unsure whether the variables that were not selected should have been used instead. Sensitivity analysis is performed such that the conclusions that were drawn are well founded and, accordingly, empirically justified. For example, the variables 10, 11, 12 and 15 of the ‘income’ category are interchangeably included in new models, instead of the variables 13 and 14 of the ‘income’ category. Another example of the performed sensitivity analysis is that the variable ‘non-Western immigrants’ is included in the model, instead of the variable ‘low income’. A final example is the inclusion of the variable ‘households with children’, instead of the variable ‘households without children’.

In conclusion, it is found that the most important findings do not change when correlated variables are interchangeably used in different models.

7 Conclusion

This paper provides a better insight into municipalities' social care data by using the Arrangementenmonitor of Thönissen Management en Advies B.V. To contribute to the targeted local policy and integral approach of municipalities, this paper constructs target groups and considers all social care data within a household. The study investigates the dynamics of households that need social care and it examines which demographic characteristics explain social care consumption of specified target groups. The research is divided into three analyses and implemented using the statistical software tools R, Tableau and Stata.

The first analysis examines to what extent the constructed target groups are statistically well-defined. Based on the K-means clustering algorithm which clusters on the distinct number of sub-domains, new groups are constructed. The Silhouette method finds that the optimal number of K clusters is equal to 24. Thereafter, both groups are compared using the Adjusted Rand Index. The corresponding ARI score shows that the constructed target groups are statistically well-defined when taking into account the combination of the different distinct sub-domains. Therefore, these target groups are used for the second and third analysis.

The second analysis explores the dynamics of the target groups. In more detail, it investigates the patterns of the in-, out- and through-flow of the target groups. This paper finds that there is a relatively high increase in the number of distinct sub-domains for the following three target groups: households needing two distinct sub-domains with the age category 'youth', households needing financial assistance with the age category 'adults', and households needing three distinct sub-domains with the age category 'elderly'. The current research has explored, for the first time in literature, that municipalities should pay more attention to these households.

The third analysis examines which demographic characteristics significantly explain the target groups. The research conducted was able to explain the target groups using demographic characteristics of a neighbourhood and some interesting new insights were found. For example, households with a high income significantly explain youth mental health and households with a low income significantly explain almost all target groups. Furthermore, significant differences between target groups were found for the variable 'married'.

This study does not explain the rationale behind the results because this is within the expertise of the municipalities. They have to choose on which target groups they want to focus and they can adjust their policy accordingly. The results of the third analysis presented in Table 7 can be used to enhance their integral and targeted neighbourhood approach.

8 Implications future research

Municipalities are still faced with a major challenge to make the decentralisation successful, both financially and in terms of the well-being of the people. This study is relevant in practice since municipalities can use the results of the dynamics between neighbourhoods and focus on the demographic characteristics they find interesting to enhance their integral and targeted neighbourhood approach. In addition, this research has developed an automatic data cleaning procedure for the 26 municipalities of the Arrangementenmonitor. Henceforth, the problem of double age categories, provider names and double neighbourhood names are automatically solved. However, since this study is one of the first that investigates the social care consumption of the Social Domain, it gives several implications and recommendations for future research.

It is recommended to repeat the study, including the data of people that do not need social care. The Arrangementenmonitor only has data of people that need social care facilities. Therefore, nothing is known about people that do not need social care facilities and nothing is known about the construction of households. A further study could assess whether these two factors further explain social care facilities.

For future research, it is interesting to construct the target groups by considering the number of providers in a household, instead of the number of distinct sub-domains in a household. This contributes to a different targeted policy. It is up to municipalities which of the two targeted policies they prefer.

The findings of this study have a number of practical implications. Further research should focus on determining whether it is possible to reduce the through-flow from the problematic households mentioned in the second analysis. These findings suggest several courses of action for the municipalities and they have to consider whether and how they are going to adapt their approach on these three target groups.

Improvements could be made in determining the reasons for the differences of the positive and negative influence that the percentage of married people has on the target groups. A further study, with a stronger focus on this variable, is therefore suggested. In addition, further research should be undertaken to investigate whether the following hypothesis is supported: households with a high income, that need youth mental health services, are more often referred to a specialist than other households.

This study shows that five target groups with the age category ‘elderly’ are significantly explained by the oldest group of elderly people, being ‘85+’. Two possible explanations were given by Thönissen Management en Advies B.V. that could have caused bias in these results. Therefore, for future research, it would be interesting to investigate the age categories 67-74 and 75-84.

Further investigation and experimentation into the level of aggregation is strongly recommended. This research is conducted on neighbourhood level. It is preferred to conduct the research on household level because only then one can be sure that the correct conclusions are drawn. For example, it is possible that, when conducting the research on neighbourhood level, people that do not need social care have demographic characteristics that incorrectly explain the target groups. However, demographic characteristics on household level are not publicly available. At the moment, Maastricht University is doing a similar research on household level and it is interesting to observe whether the results of this study match the current one.

List of Tables

1 Categorisation of social care by the Arrangementenmonitor. 6
2 The demographic characteristics per category represented as percentages on neighbourhood level. 8
3 The target groups. 10
4 Results of the Elbow- and Silhouette method. 14
5 The in-, out- and through-flow in the fourth quarter of 2018. 16
6 The correlations between the variables. 20
7 An overview of the results of the third analysis: Which demographic characteristics explain the target groups? 24

List of Figures

1 The Elbow method. 14
2 The Silhouette method. 14
3 The through-flow between target groups. 17
4 A closer look at the through-flow of target group 19. 18

Glossary

daytime activities SSA het hoofdarrangement ’dagbesteding Wmo’. 6

debt counseling het hoofdarrangement ’schuldhulpverlening’. 6

domains pijlers. 3

financial assistance het hoofdarrangement ’bijstand’. 6

guidance SSA het hoofdarrangement ’begeleiding Wmo’. 6

guidance/daytime activities youth een samenvoeging van de twee hoofdarrangementen ’begeleiding jeugd’ en ’dagbesteding jeugd’. 6

integral approach het integrale beleid: alle sociale ondersteuning in een huishouden wordt als één geheel gezien in plaats van dat de pijlers apart worden bestudeerd. 3

living SSA het hoofdarrangement ’Wmo wonen’. 6

living youth het hoofdarrangement ’wonen jeugd’. 6

neighbourhood approach de gerichte wijkaanpak, het beleid op wijkniveau. 2

participation de pijler ’participatie’. 5

remaining youth facilities het hoofdarrangement ’jeugd overig’. 6

residential facilities een samenvoeging van de twee hoofdarrangementen ’hulp bij huishouden’ en ’woonvoorziening’. 6

social care facilities sociale ondersteuning. 1

Social Support Act (SSA) de pijler ’Wmo’. De afkorting staat voor Wet Maatschappelijke Ondersteuning. 5

special education het hoofdarrangement ’speciaal onderwijs’. 6

special financial assistance het hoofdarrangement ’bijzondere bijstand’. 6

student transport het hoofdarrangement ’participatietraject’. 6

student transport het hoofdarrangement ’leerlingvervoer’. 6

sub-domain hoofdarrangement. 3

target groups doelgroepprofielen. 1

transport SSA het hoofdarrangement ’Wmo vervoer’. 6

youth care de pijler ’jeugd’. 5

youth mental health het hoofdarrangement ’jeugd-GGZ’. 6

References

Allen, M. P. (1997). The problem of multicollinearity. *Understanding regression analysis*, pages 176–180.

Anderberg, M. R. (2014). *Cluster analysis for applications: probability and mathematical statistics: a series of monographs and textbooks*, volume 19. Academic press.

Baltagi, B. (2008). *Econometric analysis of panel data*. John Wiley & Sons.

Batterink, M., Lapajian, I., and Meijer, J. (2018). Regionale verschillen in het gebruik van jeugdhulp met verblijf.

Benesty, J., Chen, J., Huang, Y., and Cohen, I. (2009). Pearson correlation coefficient. In *Noise reduction in speech processing*, pages 1–4. Springer.

Bholowalia, P. and Kumar, A. (2014). Ebk-means: A clustering technique based on elbow method and k-means in wsn. *International Journal of Computer Applications*, 105(9).

Bishop, C. M. et al. (1995). *Neural networks for pattern recognition*. Oxford university press.

Cameron, A. and Trivedi, P. (2005). *Microeconometrics: Methods and Applications*. Cambridge University Press.

Cameron, A. and Trivedi, P. (2010). *Microeconometrics Using Stata, Revised Edition*. Stata Press.

CBS (2009). Gezonde levensverwachting korter bij lage inkomens. *CBS Website*.

CBS (2020, accessed 16.05.2020). Kerncijfers wijken en buurten 2004-2019. `https://www.cbs.nl/nl-nl/reeksen/kerncijfers-wijken-en-buurten-2004-2019#id=kerncijfers-wijken-en-buurten-2019-0`.

CBS Statline (2018, accessed 26.05.2020). Gebruik voorzieningen sociaal domein; aantal voorzieningen, wijken, 2018. `https://opendata.cbs.nl/statline/#/CBS/nl/dataset/84420NED/table?ts=1592579220210`.

Cheeseman, P. C., Stutz, J. C., et al. (1996). Bayesian classification (autoclass): theory and results. *Advances in knowledge discovery and data mining*, 180:153–180.

Collins, L. M. and Lanza, S. T. (2009). *Latent class and latent transition analysis: With applications in the social, behavioral, and health sciences*, volume 718. John Wiley & Sons.

de Crombrugghe, D. (2020). Course: E-metric method cross-sect. + panel data (ebc4006). *Lecture 2: Linear panel data models without heterogeneity: Pooled LS and IV methods*, (2):64.

Dutch Child Center (2019, accessed 21.05.2020). Onderzoek toont aan: forse tekorten jeugdzorg bij gemeenten. `https://www.dutchchildcenter.nl/ouders/onderzoek-toont-aan-forse-tekorten-jeugdzorg-bij-gemeenten/`.

Elissen, A. and Ruwaard, D. (2014). *Kenmerken van individuen als voorspellers van zorgvraagzwaarte op populatieniveau: een verkennend onderzoek*.

Engbersen, R., Uyterlinde, M., and Bronsvoort, I. (2018). Verschillen geduid. exploratief onderzoek naar hoog en laag voorzieningengebruik in stedelijke regio's.

Forgy, E. W. (1965). Cluster analysis of multivariate data: efficiency versus interpretability of classifications. *biometrics*, 21:768–769.

Fränti, P. and Sieranoja, S. (2018). K-means properties on six clustering benchmark datasets. *Applied Intelligence,*, (48):4743–4759.

Fukunaga, K. (2013). *Introduction to statistical pattern recognition*. Elsevier.

GGZ Totaal (2019, accessed 21.05.2020). Onderzoek wijst uit: gemeenten krijgen inderdaad te weinig geld voor jeugdzorg. `https://www.ggztotaal.nl/nw-29166-7-3718615/nieuws/onderzoek_wijst_uit_gemeenten_krijgen_inderdaad_te_weinig_geld_voor_jeugdzorg.html`.

Greene, W. (2011). *Econometric Analysis*. Pearson Education.

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

Hartigan, J. A. (1975). *Clustering algorithms*. John Wiley & Sons, Inc.

Hausman, J. A. (1978). Specification tests in econometrics. *Econometrica: Journal of the econometric society*, pages 1251–1271.

Jain, A. K. and Dubes, R. C. (1988). *Algorithms for clustering data*. Prentice-Hall, Inc.

Kaufman, L. and Rousseeuw, P. J. (1990). *Finding Groups in Data: An Introduction to Cluster Analysis*. Wiley Series in Probability and Statistics.

Kieskamp, W. (2019). Flinke tekorten voor jeugdzorg bij gemeenten. *Trouw*.

Koster, Y. (2019). Stroppenpot heeft weinig soelaas geboden. *Binnenlands Bestuur*.

KPMG (2020, accessed 20.05.2020). Analyse op de bedrijfsvoering en maatregelen voor het sociaal domein. `https://gemeentemaastricht.nl/sites/default/files/2020-02/Rapport%20onderzoek%20sociaal%20domein.pdf`.

MacQueen, J. et al. (1967). Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, volume 1, pages 281–297. Oakland, CA, USA.

Marutho, D., Handaka, S. H., Wijaya, E., et al. (2018). The determination of cluster number at k-mean using elbow method and purity evaluation on headline news. In *2018 International Seminar on Application for Technology of Information and Communication*, pages 533–538. IEEE.

Meilă, M. and Heckerman, D. (1998). An experimental comparison of several clustering and initialization methods. In *Proceedings of the Fourteenth conference on Uncertainty in artificial intelligence*, pages 386–395.

Mundlak, Y. (1978). On the pooling of time series and cross section data. *Econometrica: journal of the Econometric Society*, pages 69–85.

Nijendaal, G. V. (2014). *Drie decentralisaties in het sociale domein.* Jaarboek Overheidsfinanciën 2014.

Ooms, I., Sadiraj, K., and Pommer, E. (2017). Regionale verschillen in het sociaal domein: voorzieningengebruik nader verklaard. *Sociaal en Cultureel Planbureau*, (24):1–33.

Pat Hanrahan, Christian Chabot, Chris Stolte, Andrew Beers (2020). *Tableau: Business intelligence and analytics software.* Salesforce, California, United States of America.

Pollard, K. S. and Van Der Laan, M. J. (2002). A method to identify significant clusters in gene expression data.

Pommer, E. and Boelhouwer, J. (2017). Overall rapportage sociaal domein 2016.

R Core Team (2020). *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria.

Rand, W. M. (1971). Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical association*, 66(336):846–850.

ROB (2017). Wegwijzer financiën in 3d: hoe zit het eigenlijk?

Rousseeuw, P. J. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, 20:53–65.

Ruiz, I. R., Valenzuela, O., Rojas, F., and Ortuño, F. (2019). *Bioinformatics and Biomedical Engineering.* Springer International Publishing.

Santos, J. M. and Embrechts, M. (2009). On the use of the adjusted rand index as a metric for evaluating supervised classification. In *International conference on artificial neural networks*, pages 175–184. Springer.

Schellingerhout, R., Ooms, I., Eggink, E., and Boelhouwer, J. (2020). Jeugdhulp in de wijk. *Sociaal en Cultureel Planbureau*, (1):1–53.

Scrucca, L., Fop, M., Murphy, T. B., and Raftery, A. E. (2016). mclust 5: clustering, classification and density estimation using Gaussian finite mixture models. *The R Journal*, 8(1):289–317.

Senate ('s-Gravenhage, 2014). *Regels inzake de gemeentelijke ondersteuning op het gebied van zelfredzaamheid, participatie, beschermd wonen en opvang (Wet maatschappelijke ondersteuning 2015).* ISSN 0921 - 7371.

Smeets, R. G., Elissen, A. M., Kroese, M. E., Hameleers, N., and Ruwaard, D. (2020). Identifying subgroups of high-need, high-cost, chronically ill patients in primary care: A latent class analysis. *PloS one*, 15(1):e0228103.

Steinley, D. (2004). Properties of the hubert-arable adjusted rand index. *Psychological methods*, 9(3):386.

Subbalakshmi, C., Krishna, G. R., Rao, S. K. M., and Rao, P. V. (2015). A method to find optimum number of clusters based on fuzzy silhouette on dynamic data set. *Procedia Computer Science*, 46:346–353.

Syakur, M., Khotimah, B., Rochman, E., and Satoto, B. (2018). Integration k-means clustering method and elbow method for identification of the best customer profile cluster. In *IOP Conference Series: Materials Science and Engineering*, volume 336, page 012017. IOP Publishing.

Teeven, F. and van Rijn, M. J. (2014a). *Jeugdwet.*

Teeven, F. and van Rijn, M. J. (2014b). *Memorie van toelichting bij de jeugdweg.*

Teeven, F. and van Rijn, M. J. (2014c). *Memorie van toelichting Wetsvoorstel Maatschappelijke Ondersteuning 2015.*

Thönissen, M. (2016). Arrangementenmonitor sociaal domein. pages 1–16.

Tou, J. T. and Gonzalez, R. C. (1974). Pattern recognition principles.

Verbeek, A. L. N. V. J. J. (2002). The global k-means clustering algorithm. *Pattern Recognition*, (2).

Vinh, N. X., Epps, J., and Bailey, J. (2010). Information theoretic measures for clusterings comparison: Variants, properties, normalization and correction for chance. *The Journal of Machine Learning Research*, 11:2837–2854.

Vuik, S., Mayer, E., and Darzi, A. (2016). Patient segmentation analysis offers significant benefits for integrated care and support. *Health affairs*, 35(5):769–775.

Warrens, M. J. (2008). On the equivalence of cohen's kappa and the hubert-arabie adjusted rand index. *Journal of classification*, 25(2):177–183.

Weisstein, E. W. (2004). Bonferroni correction. *https://mathworld. wolfram. com/.*

Westra, D., Gerritsma, J., Hameleers, N., Jansen, M., and Ruwaard, D. (2018). Een studie naar het gebruik van jeugd-ggz in zuid-limburg. *Maastricht University, Faculty of Health, Medicine and Life Sciences*, pages 11–99.

William Gould (2020). *Stata: Software for Statistics and Data Science.* StataCorp, Texas, United States of America.

Witten, I. H. and Frank, E. (2002). Data mining: practical machine learning tools and techniques with java implementations. *Acm Sigmod Record*, 31(1):76–77.

Wong, H. J. and Hartigan, J. (1979). Algorithm as 136: A k-means clustering algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 28(1):100–108.

Wooldridge, J. (2010). *Econometric analysis of cross section and panel data*, volume 2. MIT Press.

Yeung, K. Y. and Ruzzo, W. L. (2001). Details of the adjusted rand index and clustering algorithms, supplement to the paper an empirical study on principal component analysis for clustering gene expression data. *Bioinformatics*, 17(9):763–774.