Commuters, the tip of the iceberg in solving Zipf's law?

Student: Thomas de Ruiter

Supervisor: Henri L.F. de Groot

Vrije Universiteit Amsterdam

Master thesis

Abstract

Calculating the rank sized coefficient for municipalities Netherlands shows that the distribution of city sizes is a lot more equal than would be expected from the literature. Zipf's law literature shows that most estimations at least loosely approach the rank size rule distribution, this makes the estimation for the Netherlands suspicious. Is the Dutch urban system really that equally divided, or do the municipalities sketch a misleading picture? This paper shows how municipalities underestimate the regional size inequality, compares several alternative methods of estimating the rank sized coefficient and proposes a model of its own based on commuter data. This model shows that the Dutch urban size distribution is actually a lot less equal than would be expected from the municipality estimation. In fact, the coefficient is about –1, therefore the rank sized rule predicted from Zipf's law literature holds. It further shows that while the commuter flows are related to jobs, the rank sized coefficient for jobs is significantly lower. This is evidence for Zipf's law not following from the distribution of jobs but from urban populations and dependents. The paper innovates on the literature by abandoning geographical borders and switching from determining functional areas to functional population.

Keywords: Zipf's law, Rank size rule, commuters, urban networks

Commuters, the missing puzzle piece in Zipf's law?

When ordering the most populous cities of the United States by population size, a funny regularity occurs. Regressing the natural logarithm of the population size on the natural logarithm of the rank gives a near perfect slope of –1 (Gabaix, 1999). This means that the largest city is approximately twice as large as the second city, thrice as large as the third city and n times as large as the nth city. This regularity is called the rank sized rule and it occurs in many other countries than the United States. One of the countries where the rank sized rule does not hold is the Netherlands. De Groot et al. report a coefficient of –0.64 for Dutch municipalities in 2009, they use the Lotka specification in which ln(population) is taken as the dependent variable. This shows a more evenly distributed urban system than would be expected through the rank sized rule. De Groot et al. note that using municipalities as observations result in a more equal distribution because some municipalities can reasonably be included in larger cities instead of counted separately (De Groot et al., 2010). This bias is an example of how conclusions on any topic involving geographical units are dependent on the definition of those units. The way a geographical area is defined has significant implications for results. Frequently used specifications are standard administrative units like municipalities, metropolitan areas or so-called natural cities. This paper compares different urban definitions, their advantages, disadvantages and their rank size coefficients to find the most accurate way to define urban areas. The paper will then introduce a new model to define urban networks based on commuters and the population they generate in their place of living. The paper is structured as follows: in section 2 a literature review will be conducted. In section 3 several alternative urban definitions will be discussed. In section 4, the commuter model is introduced and compared to the distribution of jobs. In section 5, technical details of the model are discussed, and the correct specification is estimated. Section 6 shows and discusses the results from the commuter model to form a new city definition. The final section concludes the study.

2 – Literature review

To give a better context on the analysis conducted later in this paper let us first shortly introduce Zipf's law for cities, the rank sized rule and their challenges. Gabaix (Gabaix, 1999) states Zipf's law as $P(\tilde{S} > S) = aS^{-\zeta}$. With $\tilde{S}$ being the size of a city, S being a range of sizes in the upper tail of the rank size distribution and ζ being the power law exponent. A ζ of 1 corresponds to Zipf's law. This probabilistic statement of Zipf's law can be intuitively explained as "that the probability that a city has a size greater than *S* decreases as *1/S*" (Gabaix, 1999). The rank sized rule is a sister of Zipf's law. Jiang and Jia state it as $s = r^{-1}$ (Jiang & Jia, 2011). *s* in this case represents a fraction of the size of the largest city. The exponent of –1 corresponds to the rank sized rule, a different exponent corresponds to a non-Zipf power law resulting in a more equal or less equal distribution. If Zipf's law holds, the rank sized rule holds only approximately and vice versa (Gabaix, 1999). This allows us to test whether Zipf's law holds based on a linear regression testing the rank sized rule. This is done using the least squares regression: $\ln(rank) = \beta_0 + \beta_1 \ln(size) + \varepsilon$. This statement is known as the Pareto specification and is favoured by Gabaix and Ioannides (Gabaix & Ioannides, 2004). Alternatively, the Lotka specification is used in which the log of size is taken as the dependent variable and the log of rank as the independent. This paper makes use of the Lotka specification as it follows earlier Dutch literature on Zipf's law that made use of the Lotka specification and because the interpretation feels more intuitive. Appendix 1 further explores the difference between the Pareto and Lotka specifications.

Several explanations have been put forward in explaining the emergence of Zipf's law. In their introductory book to the world of geographical and urban economics, Brakman et al. describe a certain set of conditions and restrictions related to congestion through which the NEG model introduced by Krugman (Krugman, 1991) could lead to Zipf's law (Brakman et al., 2020). This explanation heavily relies on very specific values of congestion and transport costs. It seems improbable that such a regionally differing phenomenon could result in such a universal law. (Liu & Liu, 2009) propose combining the

central place theory from Walter Christaller (Brush, 1966) with fractal mathematics to end up in a rank sized system. The problem with their theory is that it again requires a very specific set of conditions to result in the power laws observed around the world. One of these conditions, the nature of intercity relations from the central place theory, is determined by assuming a rank size distribution and utilising the model to find the nature of the intercity relation. Their model therefore does not show how the rank sized rule emerges as a result. This explanation is therefore unconvincing until tested using more objective measures on the nature of intercity relations. (Gabaix, 1999) considers two more previously published urban models attempting to reach a Zipf's distribution. That of (Simon, 1955) and (Steindl, 1965). Both these authors model an urban system evolving over time. Unfortunately, they both require a city birth rate that is higher than the city growth rate, which does not accurately describe developed urban systems holding Zipf's law today. Gabaix proposes a more convincing first step towards an explanation. He later reinforces this first step in his 2004 paper with Ioannides (Gabaix & Ioannides, 2004). He shows how an urban system following Gibrat's law would necessarily end up in a Zipf's law steady state distribution. Gibrat's law states that the growth rate and standard deviation are independent from population. Past success is not indicative of future success in the long term. By making this first step, Gabaix makes further explanations for Zipf's law more feasible.

According to Volker Nitsch (Nitsch, 2005) in his meta-analysis of coefficients found in Zipf's law literature up to 2002, the commonly understood loose range for which Zipf's law can reasonably hold is between –0.8 and –1.2. Gabaix and Ioannides place this range between –0.85 and –1.15. From the meta-analysis Nitsch however finds a mean coefficient of –1.1 in the pareto specification. This means that the urban networks are generally more evenly distributed than would be expected from a strict adherence to Zipf's law. For estimates making use of metropolitan areas however, this value drops to around –1, which supports a strict adherence to Zipf's law for metropolitan areas. This shows that the choice in urban units has a significant impact on the eventual outcome. This also raises questions about

the difference in characteristics between metropolitan areas and standard administrative units like municipalities that cause this difference. This seems to suggest that an approach utilizing standard administrative units underestimates the true size of urban areas while it possibly overestimates smaller settlements and rural areas. The difference in outcomes between standard administrative units and metropolitan areas begs the question of what a proper way would be to count urban units.

(Jiang & Jia, 2011) use an alternative to administrative borders in estimating Zipf's law. They use open street map data to compute natural cities. They do this by selecting a street node (intersections and ends) drawing a radius of a certain distance around this node, selecting the nodes within this radius, drawing new circles and repeating this process until no more circles can be drawn. The resulting entity is labelled as a natural city. They then estimated the coefficient from the number of nodes within each natural city and the surface area covered by these natural cities for different radiuses. Their approach returns coefficients close to –1. The assertion that the number of nodes or the surface area is a good proxy for urban processes is doubtful however. If it would have been possible to determine the population of these natural cities, an estimation using it would return a distribution that is a lot more unequal than found by Jiang and Jia. This is due to the enormous differences in population density between smaller and larger natural cities. It is also questionable if continuous built-up area is the proper way to define an urban area. Urban markets and processes like labour and amenities do not require the population they attract to live in a place that is connected to the city by a continuous built-up stretch. Such a method would also be difficult to implement in the Netherlands since the street node density is much higher than the United States. It would be difficult to find a radius that would not count different riversides within the same city separately yet refrains from counting the Randstad as one big city. Natural cities based on population densities on a raster map would be more suitable, they would still not cover the true extent of urban processes however.

There are some known notable exceptions to Zipf's law. (Gabaix, 1999) names two. He notes that the capital city in many countries is larger than would be expected from Zipf's law. This has to do with many additional jobs and agglomeration benefits that are associated with the political and legislative heart of a country. This is not the case in the Netherlands since while the largest city is the capital, the political and legislative heart is not located in the capital. The other exception he theorizes is that because medium sized cities do not follow Gibrat's law as well as cities higher in the distribution, medium to small sized cities would be underrepresented in the rank size distribution. This last finding is questioned by (Nitsch, 2005) who finds that contrary to larger sample sizes (more medium to small cities) leading to smaller estimates (more unequal distributions in this case), it is actually smaller sample sizes leading to these unequal distributions. This paper proposes a third deviation from Zipf's law stating that heavily interconnected urban networks lead to an underestimation of the rank sized coefficient in the Lotka specification.

3 – Alternative specifications

In this section the Zipf's coefficient is estimated from several different urban definitions. These different definitions will be discussed along with their advantages and disadvantages in the estimation of Zipf's law for the Netherlands. Lastly, Zipf's coefficient will be calculated for each of these systems. The resulting coefficients can be found in table 1.

Municipalities

A very natural place to start trying different urban definitions to see if Zipf's law applies to the Netherlands is to start by looking at the most commonly used definition, being municipalities. Municipalities are administrative units, the most approachable and locally oriented layer of government. This results in them usually consisting of a central settlement and its surrounding area. More rural municipalities are often multipolar and represent an area of service and policy without a more 'spiritual' centre. This results in the rank sized coefficient actually displaying the distribution of these service areas instead of all settlements in the country. An approach that does make use of all settlements is the natural city approach, which is used in (Jiang & Jia, 2011) for example. Compared to the natural city approach, municipalities are expected to show a more equal distribution. This is because municipalities group smaller settlements unable to provide sufficient services and effectively govern on their own into larger joint municipalities while leaving larger settlements unchanged. The idea of basing Zipf's law on anything other than settlement population is valid however. A larger collection of settlements can more holistically be considered a network sharing services, markets, amenities and more. Such a collection can be considered as an urban network or city all of its own. The rank sized rule can then be tested from these larger networks. Municipalities are however also not fully indicative of such networks. Some municipalities are largely dependent on other neighbouring municipalities for services, markets and amenities and are thus themselves part of larger urban networks. Municipalities not fully encompassing

urban networks but grouping the lower tail of the settlement size ranking leads to them being biased towards a more equal rank sized coefficient.

When plotting the rank sized rule regression for municipalities (see figure 1) it becomes apparent that the lower tail of the distribution deviates from the power law leading it to be biased towards a more unequal distribution. (Gabaix, 1999) Shows that the power law distribution is contained to the upper tail of the rank size distribution, therefore the lower tail should be discarded. (Nitsch, 2005) questions this finding and shows how datasets are too small generate estimations representing a more unequal distribution. The debate on what observations should be included is not completely settled, it does seem likely however that at a certain size sizes start deviating from the power law. This effect is present in the municipality data. To isolate the slope of the power law the dataset is limited to the top 250 municipalities (see figure 2). In doing this, Nitsch' line is followed in including as many observations as possible until the distribution starts visible diverting from the power law distribution. This results in the rank sized coefficient of –0.663, which is indeed significantly smaller than the –1 expected from a true Zipf's law distribution. It is also significantly smaller than empirical estimates from the literature. (Nitsch, 2005) Concludes that the average estimation of rank size coefficients is about –1.1 in the Pareto specification. This is a slightly more equal distribution than predicted from Zipf's law. As he conducts his meta-analysis based on the Pareto specification, he takes the inverse of estimates using the Lotka specification. This is not entirely econometrically sound, see appendix 1. It does however allow us to easily compare the rank sized coefficient resulting from municipalities with the literature. The inverse of the municipality rank sized coefficient is about –1.5, significantly higher than the –1.1 found by Nitsch.

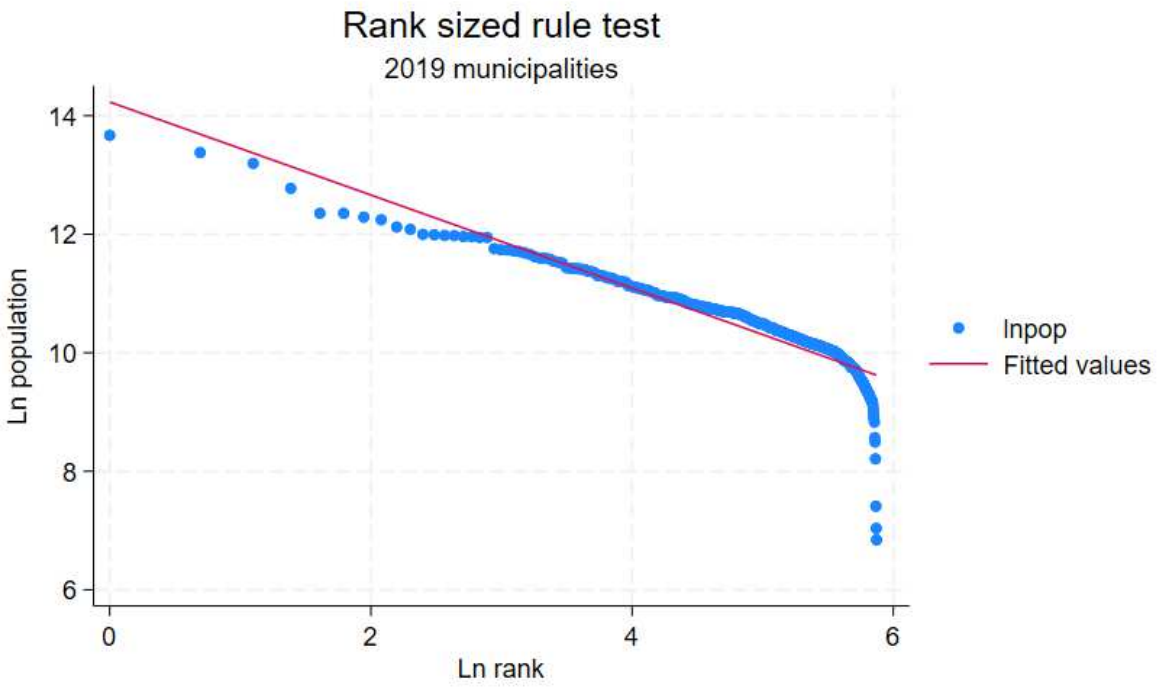

Figure 1 – Rank sized rule test municipalities 2019 – Slope –0.785

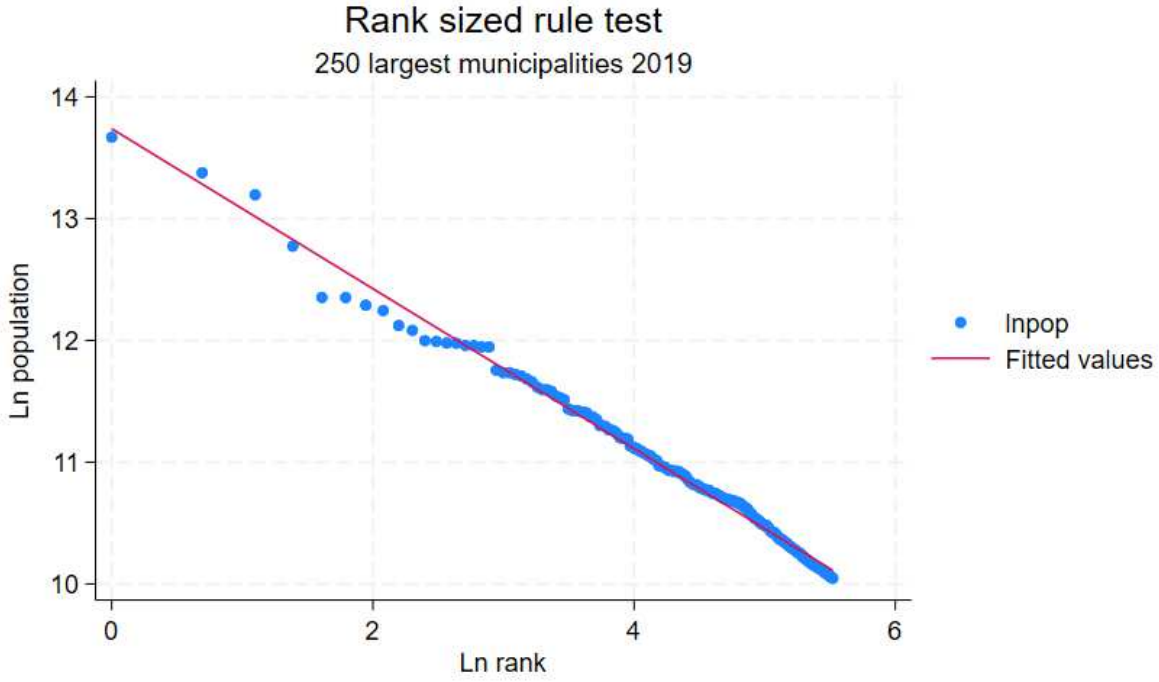

Figure 2 – Rank sized rule test 250 largest municipalities 2019 – Slope –0.657

GSA's and Stadsgewesten

An example of an urban system that includes nearby municipalities displaying suburban functions and therefore more accurately describes urban networks are the GSA's (greater city agglomerations) and the Stadsgewesten. The Stadsgewesten include the GSA's and some municipalities surrounding them based on the number of commuters and the relatedness of the housing market. These were discontinued as administrative units in 2015 which led to some problems in computing their population since some included municipalities had merged with excluded municipalities in the meantime. In cases where this occurred the municipality was included or excluded based on what portion of the new municipality was situated within the GSA/stadsgewest and whether the major population centre was situated within the GSA/stadsgewest. The major advantage of using the GSA's and stadsgewesten compared to the municipalities is that these administrative units much better describe the true urban areas and networks and will therefore lack the underestimation of larger municipalities and the overestimation of smaller municipalities found with using municipalities when testing Zipf's law. Major disadvantages of this approach are that there are only 22 GSA's and stadsgewesten, which means that only the very upper end of the distribution can be estimated. Furthermore, the map of GSA's and stadsgewesten does not cover the entire country. It only covers the direct surroundings of 22 major cities. Therefore, it is contentious to claim that the outcome of a rank sized coefficient estimation is indicative of the urban distribution of the entire country. The coefficients found by the GSA's and the stadsgewesten are –0.786 and –0.756 respectively. This shows that the population of the urban units following this approach is significantly more unevenly divided compared to using municipalities only. It does not follow the power law with slope one expected from Zipf's law however.

De Nieuwe Gemeentekaart (Marlet & Van Woerkens, 2014)

Another system that could be used is 'De Nieuwe Gemeentekaart' (the new municipality map), from Marlet and van Woerkens published in the 2014 atlas voor gemeenten. In this paper they propose a new municipal division of the country. They see such a redivision as beneficial because many markets and attraction ranges are not contained within the current municipal borders. Examples are the labour and housing market or the range in which a football stadium attracts visitors. Consolidating the service areas of cities into single municipalities allows the municipalities to effectively make policy, tax and supply amenities for the entire service area. To shape their new municipalities, they determine for every pc4 (similar to neighbourhood level) area what municipality it is dependent on for work, shopping, culture, education, healthcare, sports and nature. If the majority of people within a municipality is primarily dependent on their own municipality for all these topics it is considered a core municipality and kept. The rest of the municipalities has been divided between the remaining municipalities based on which new municipality they are most dependent on. Using this process, they found that in 2014 the optimal division of the Netherlands left 57 municipalities instead of the 403 at the time.

In determining whether or not the dynamics behind Zipf's law hold for the Netherlands this definition offers some advantages and disadvantages. When compared to the standard municipalities this definition better encompasses the true area of effect for cities both in labour supply and consumer markets. This would go a long way in making the Dutch urban system comparable to larger less interconnected countries for the purposes of constructing Zipf's law. It is also preferable over the GSA's and stadsgewesten since these municipalities cover the entire country and are higher in number. A disadvantage of this approach is that it is very rigid. If a municipality is not self-dependent on any of seven topics it is added in its entirety to the core municipality it is most dependent on. This disregards that there might be a portion of its population that is more dependent on the own municipality or another core municipality. Another disadvantage of both the 57 municipalities and the

stadsgewesten/GSA's is that by not reporting any municipalities from the lower tail, the contrast between larger and smaller places becomes less stark resulting in a possible underestimation of the rank sized coefficient.

Since the research by Marlet and Van Woerkens was done in 2014, running a rank sized rule test regression would test the rank sized coefficient for their 57 municipalities in 2014. To be able to compare the results to the other specifications a column with the rank sized coefficient from the municipalities in 2014 was included. Its coefficient is very similar to that of 2019. The coefficient from the 57 municipalities is –0.793 which is significantly closer to –1 than the coefficient from the municipalities and slightly closer to –1 than the coefficient from the GSA's and stadsgewesten. This suggests that the urban rank sized distribution is a lot more unequal than the municipal estimate suggests. Volker Nitsch states that many Zipf's law estimations result in estimates of between –0.8 and –1.2 (Nitsch, 2005). The value from the 57 municipalities is still outside of this range which means the Netherlands could still reasonably be assumed as an exception to Zipf's law.

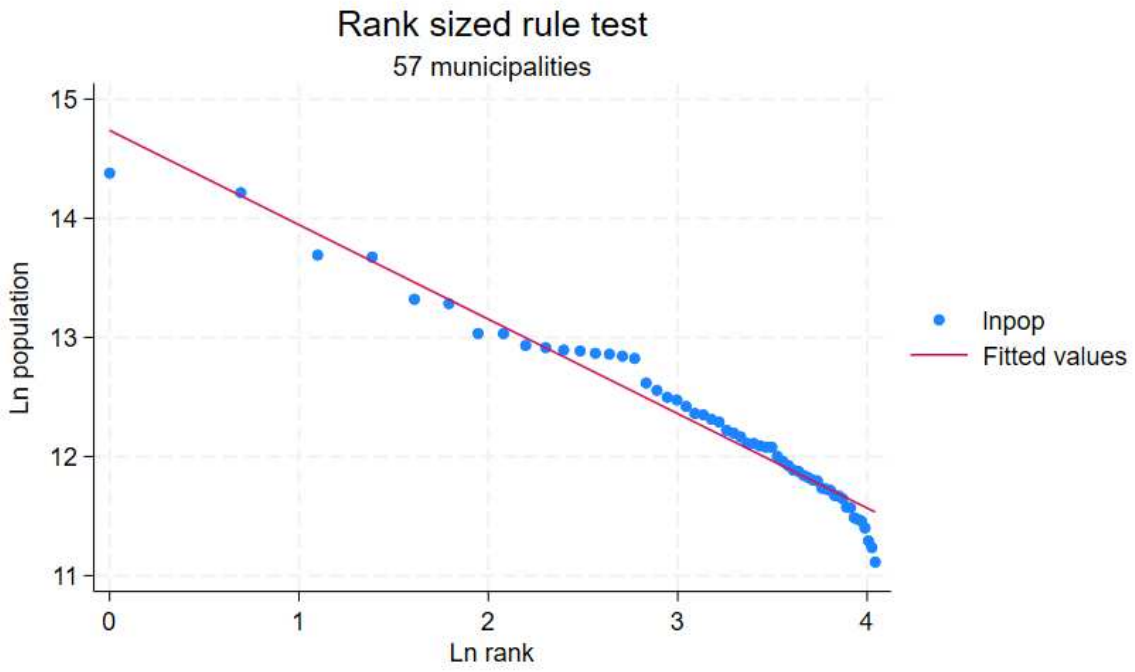

Figure 3 – Rank sized rule test 57 municipalities Marlet and van Woerkens – slope – 0.793

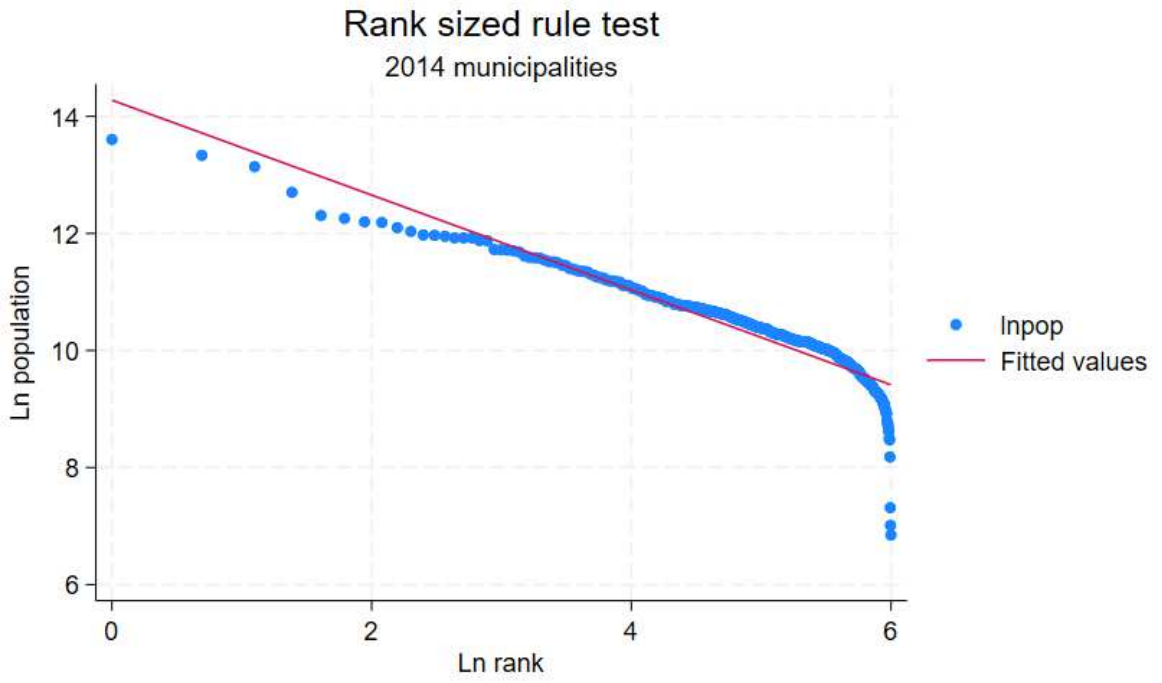

Figure 4 – rank sized rule test municipalities 2014 – Slope –0.811

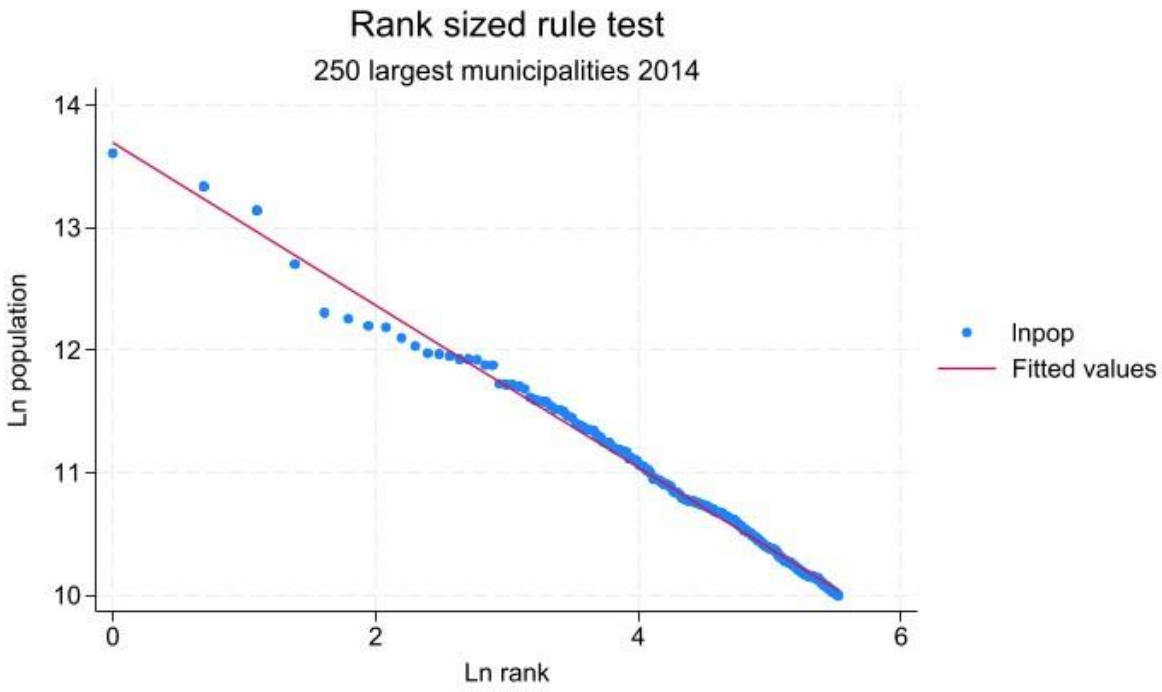

Figure 5 – rank sized rule test 250 largest municipalities 2014 – Slope –0.663

4 - Introducing the Commuter model

This thesis proposes another urban definition to test if Zipf's law holds for Dutch urban networks. This model hinges on commuters. A major difference between the Netherlands and other larger countries for which Zipf's law has been tested is the interconnectedness found in the Netherlands. Not only between commuter cities and core cities but also between core cities. To compare Dutch outcomes with international outcomes one would first have to untangle the spiderweb that is the Dutch network. To illustrate this spiderweb, a map detailing the commuter flows in the Netherlands has been taken from (De Groot et al., 2010) and shown in figure 6. The commuter model attempts to untangle this web by eliminating all commuter flows and adding the commuters to the municipality where they work along with a certain portion of population they would take with them. This extra portion consists of direct dependents like non-working partners or children as well as labour market effects through the added demand and agglomeration benefits this commuter creates at the place of living. This is similar to asking the hypothetical: where would people live if the distances between municipalities would be extremely large? This hypothetical is similar to the urban system in the United States for example, where distances between major cities are extremely large and commuting happens mostly within cities/metropolitan areas. Such a model has some major advantages when compared to the previously stated administrative units from which an estimate can be obtained. The commuter model does not rely on geography in constructing city populations. This means that the population from a municipality can be split according to the workplace of its inhabitants. This is a major improvement because it allows for a proper allocation of population between multiple destination cities while not discarding the people that do work in their own municipality. This way the lower tail of the distribution is drained to fit the actual urban networks, without completely removing the lower tail leading to the previously mentioned estimation problems. This also means that the intensity of a

regional dependency is accurately modelled. In the previously addressed approaches, every current municipality is fully added to the new administrative unit while the commuter model allows partial distribution. Another benefit of this approach is that it allows for commuting between core cities to be accounted for. The Netherlands is a truly interconnected network of cities. This means that commuting and dependency does not only occur between core cities and their feeder towns but also between core cities themselves. (Groot et al., 2012) Shows that higher educated workers commute further and more often live in areas with higher land rents. This is often between two core cities. An approach that is restricted to geographical borders is unable to account for the dependencies between major cities. It is expected that this approach will lead to a rank sized coefficient with a higher absolute value signalling a more unequal distribution. The technical details and estimations can be found in the next sections.

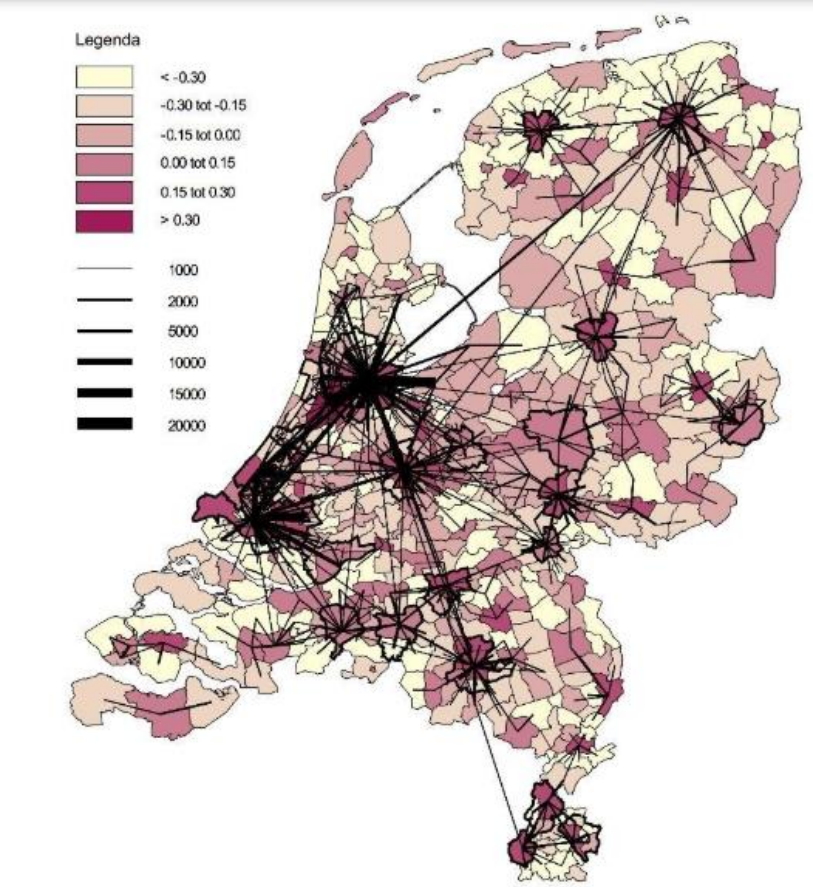

Figure 6 – commuter flows Netherlands – source

Jobs

A valuable question to ask is if the commuter model is in fact a convoluted way to follow the distribution of jobs. To analyse this question, it is important to look at the portion of the population that does not commute or work. The Netherlands numbered 17,3 million inhabitants in 2019 while the SBI(2008) counted around 8,5 million jobs at that time(CBS, 2023). The SBI(2008) does not give the complete picture of labour in the Netherlands, but it certainly is indicative that there are a lot of Dutch citizens that do not choose their place of living based on their own job on the account of not having one. Part of these are dependents, like children or non-working partners. Another part does not have to work for their living, pensioners for example. It does not seem likely that this half of the population is proportionally distributed with the jobs. Some correlation between the commuter model and the jobs is to be expected however seeing as the dependents are assumed to move alongside the working population in the commuter model. The degree of correlation is dependent on the model specification. If it is assumed that the number of dependents per commuter is the same across the country, the working population and dependents should be perfectly correlated with the number of jobs. This will not directly show in the rank size rule regression however since the population also consists of the independent non-working population that might have preferences in the amount of population in their place of living. Furthermore, it is intuitively improbable that the number of dependents and the impact on the local labour market and agglomeration benefits is constant over space. Household sizes are far larger in the countryside than in the city. The differences between the rank sized coefficients of the commuter model and jobs are valuable to consider because they give an insight into the determining factors of the rank sized rule. Differences between the commuter model and jobs rank sized coefficient under the assumption that the increase in population per commuter is constant across space can be attributed to independent non-working population. These differences therefore give an insight in the way in which the non-working population is distributed. A value of –1 for both the jobs rank sized

coefficient and a correctly estimated commuter model rank sized coefficient would indicate that the population of cities perfectly follows the jobs and thus that it is not the distribution of population that is described by Zipf's law but actually the number of jobs. Furthermore, this would describe that the true Zipf's law would hold for the job distribution in the Netherlands. It seems doubtful that this is the case. The power law coefficient was estimated using 2019 data on jobs per municipality (CBS, 2023). After correcting for the deviation in the lower tail by deleting observations outside the top 300 the power law coefficient remains at –0.871. While this value is within the confidence interval proposed by (Nitsch, 2005), it is still quite a bit removed from the theoretical –1 coefficient signifying the true Zipf's law. This comparison will be revisited in the commuter model's results section.

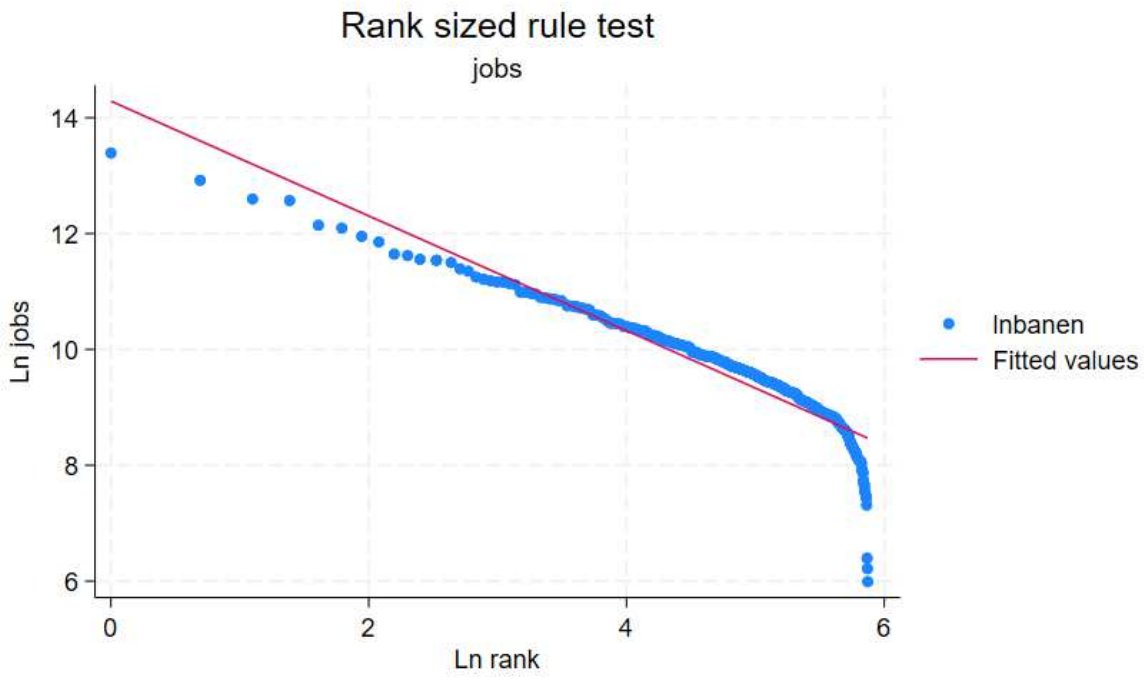

Figure 7 – rank sized rule test jobs from all municipalities – slope –0.991

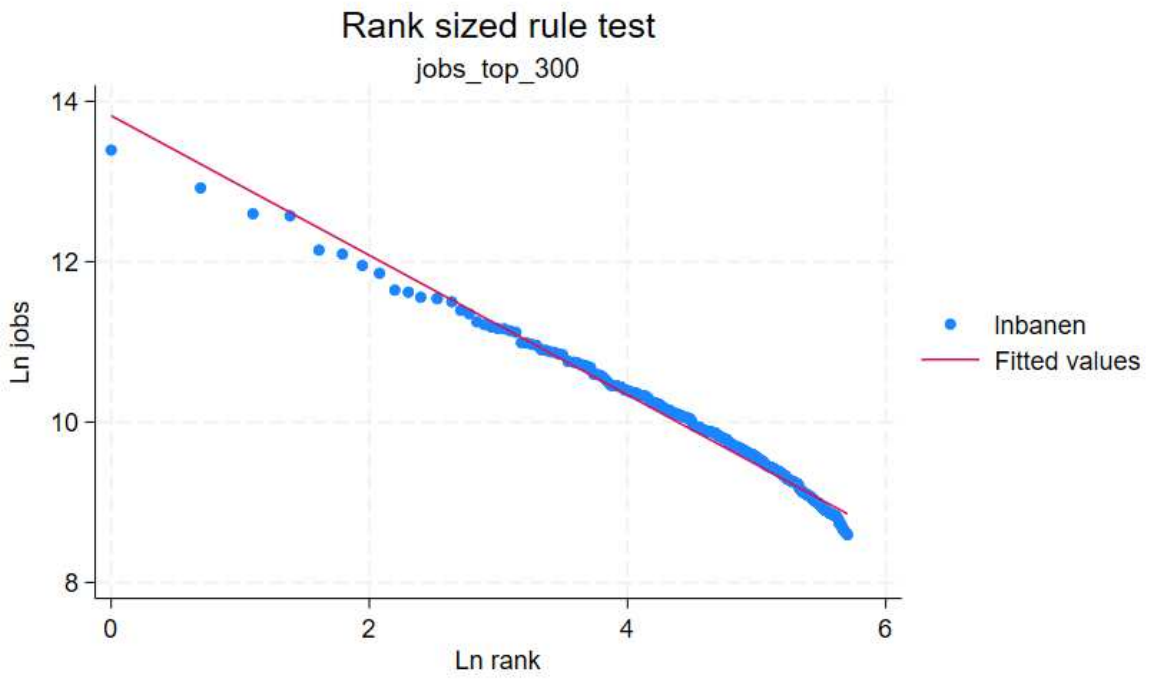

Figure 8 – rank sized rule test jobs from top 300 municipalities – slope –0.871

VARIABLES	(1) 2019 municipalities	(2) stadsgewest	(3) GSA	(4) De Nieuwe Gemeentekaart	(5) 2014 municipalities	(6) Jobs
lnrank	-0.657***	-0.756***	-0.786***	-0.793***	-0.663***	-0.871***
	(0.00742)	(0.0554)	(0.0454)	(0.0309)	(0.00756)	(0.0115)
Constant	13.74***	14.43***	14.13***	14.74***	13.70***	13.82***
	(0.0358)	(0.135)	(0.120)	(0.100)	(0.0368)	(0.0560)
Observations	250	22	22	57	250	300
R-squared	0.993	0.949	0.966	0.965	0.994	0.985

Robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Table 1 – alternate specifications regression results

5 – Estimating the Commuter model

The new municipal map proposed by Marlet and van Woerkens already showed that the Dutch urban system is strongly interconnected (Marlet & Van Woerkens, 2014). They defined new municipalities based on the degree of connection to a central city. The interconnection within the Dutch urban system is more intricate however. Some extra complications are that commuting also takes place between central cities and that central cities can share peripheral living centres. This model attempts to untangle these connections by asking a hypothetical. What would the rank sized coefficient be if the distance between cities would become infinite and people would live where they work? In this model it is assumed that people choose their place of living solely on their workplace. This is an inaccuracy seeing as (Glaeser et al., 2001) show that people are willing to pay over their urban salary benefits to live in high amenity cities. This suggests that people that work in low amenity cities might in the hypothetical scenario switch jobs to live in a high amenity city. The model therefore underestimates the increase in population for high amenity cities in the hypothetical scenario.

The model

Formula 1 of the model (shown on next page) shows an adapted form of the standard linear regression model used to determine the rank sized coefficient. Instead of using the total population size, the population size is adjusted for commuters. These commuter flows are multiplied with a coefficient that represents that commuters contribute more than just themselves to the population of their place of living. They for instance have non-working partners, they might have children and they make use of local services creating jobs in the process. Formula 2 is the second stage in a two stage least squares instrumental variable regression with formula 3 being the first. The number of outgoing commuters is an endogenous variable in determining the total population. More commuters lead to a larger population, but population sizes also have a causal effect on the number of commuters. A small village can for

instance only house a small number of commuters. It is however less likely to provide many jobs leading to a large percentage of inhabitants consisting of commuters. When the number of commuters is taken as a percentage of the total population reverse causality is still to be expected. It is not unlikely for instance that commuters have an above average household size, people with larger households will sooner choose municipalities with lower housing costs and commute than pay for larger homes at a premium. This results in the percentage of commuters having an impact on population size through larger households. Larger cities are generally more expensive to live however, resulting in lower shares of commuters. It is clear that the percentage of commuters and the total population influence each other via at least housing prices and likely more, thus calling for instrumental variables regression.

$$\ln(rank_i) = \alpha_0 + \alpha_1 \ln\left(population_i + \sum_j \beta_j Commuters_{ji} - \sum_j \beta_i Commuters_{ij}\right) \quad (1)$$

$$population_i = \beta_0 + \beta_1 Commuters_{ij} + \sum_{X=1}^{X} \beta_X X_i Commuters_{ij} \quad (2)$$

$$Commuters_{ij} = \pi_0 + \pi_1 (Neighbouring\ exogenous\ jobs_i) \quad (3)$$

i: $A\ given\ municipality$

j: $An\ other\ municipality, the\ sum\ of\ j\ means\ all\ other\ municipalities$

$rank_i$: $The\ rank\ of\ a\ municipality\ based\ on\ population\ size$

α_0, β_0, π_0: $Constants$

α_1: $Rank\ sized\ coefficient$

$population_i$: $The\ current\ municipality\ population\ size$

β: $A\ coefficient\ noting\ the\ total\ added\ population\ generated\ by\ 1\ person\ or\ 1\%\ of\ people\ in\ municipality\ i\ or\ j\ commuting$

$Commuters_{ji}$: $commuterflow\ from\ municipality\ j\ towards\ municipality\ i\ stated\ as\ an\ absolute\ or\ a\ percentage\ of\ the\ total\ population\ in\ municipality\ j$

$Commuters_{ij}$: $commuterflow\ from\ municipality\ i\ towards\ municipality\ i\ stated\ as\ an\ absolute\ or\ a\ percentage\ of\ the\ total\ population\ in\ municipality\ i$

X: $Municipality\ characteristics$

π_1: $IV\ instrument\ coefficient$

$Neighbouring\ exogenous\ jobs$: $Size\ of\ the\ job\ market\ within\ a\ radius\ surrounding\ the\ municipality$

The criteria for a valid instrument are that the instrument is exogenous and relevant. An exogenous instrument is only correlated to the dependent variable through the independent variable. A relevant (strong) instrument has a significant correlation to the independent variable. This is characterised by for example a high R^2 in an OLS regression of the instrument on the independent variable. The instrument chosen to solve the issue of endogeneity is the number of exogenous jobs in nearby municipalities. These are defined by the total number of jobs minus the jobs that directly make a region more attractive to live nearby to. Examples of such jobs are people working in retail, hospitality or government services. These jobs make neighbouring municipalities more attractive to live in thus increasing their non-commuter population. All other jobs only increase the attractiveness of surrounding municipalities through the producer side, through commuters. Thus, they are exogenous. The job groups that have been presumed exogenous and thus have been chosen as an instrument are A (agriculture), B (mining), C (industry), D (energy), E (water and waste management), F (construction), H (transport), K (financial services), L (real estate) and M (specialist business services). More details on the jobs contained by these and other categories can be found in (CBS, 2022b).

It is expected that the true increase in population from one commuter differs per municipality. People commuting from Amsterdam are unlikely to have large households since that would be very expensive in terms of housing. People commuting from smaller villages where housing prices are cheaper on the other hand are expected to have larger households. Similar stories of differing population increases per commuter can be told about the creation of jobs in local services or agglomeration advantages. To generate different population increases per commuter per municipality, interactions between the number of commuters and municipality characteristics are used. Since these characteristics are interacted with the number of commuters and the number of commuters is endogenous, the interactions are also endogenous. The instruments used to estimate these are

interactions between the exogenous jobs and the characteristics. In that way the instruments take care of the endogenous part of the interactions originating from the number of commuters.

The data

To estimate this model, data from the Central Bureau of Statistics Netherlands was used. For the commuter flows the "Banen van werknemers naar woon- en werkregio (2014-2020)" dataset from December 2018 was used (CBS, 2022a). This dataset reports the number of people living in a municipality and working in a municipality for every possible combination of municipalities including people living and working in the same municipality. To only report commuter flows the values for instances of the living- and working municipality being the same were all set to zero. To compute an instrument the CBS dataset: "Banen van werknemers in December; economische activiteit (SBI2008), regio" was used (CBS, 2023). Again, the data from December 2018 was used. All the municipal characteristics and shapefiles were taken from the 2019 Wijk- en buurtkaart (CBS & Kadaster, 2021) which reports municipal data from January. On the first of January 2019, changes were made to the Dutch municipal structure. A number of smaller municipalities merged to reduce the total number of municipalities to 355. To even out the data the commuter flows to and from municipalities that got merged were added together and the internal commuter flows were set to zero. The exogenous jobs were summed over the merged municipalities. One scarcely populated municipality was split between two new municipalities that emerged from mergers. The exogenous jobs and commuter flows were attributed to the municipality receiving the majority of the split municipality. When summing the exogenous jobs within certain radiuses of municipalities a geometrical error in the dataset resulted in the population of a small municipality in the southeast of Brabant not being counted. Since the municipality barely contained more than 1000 exogenous jobs, this omission should result in very little bias.

Calibrating

To get the most valid estimate of the power law coefficient several specifications were tried. The first attempts regress the percentage of people within a municipality that commute against the population size of that municipality. An OLS regression of the exogenous jobs on the percentage of commuters shows the highest R2 for a buffer of 30 km's. This R^2 is quite low however, only 15%. This suggests that the instrument is not very strong and that the results of the IV regression might be biased towards the OLS result. A possible explanation for this low R^2 besides the number of neighbouring jobs and the number of commuters just not correlating that strongly is that there are some gaps in the data.

For example, the specialist business services from the large municipalities of Arnhem and Groningen are not reported which results in a larger error in the OLS regression. Running a regression with many interactions gave large standard errors and showed many regressors to be highly statistically insignificant. To improve the reliability of the estimations regressors were incrementally removed based on their statistical significance and kept if removing them resulted in large shifts in coefficients of other regressors or in large increases in standard errors for other regressors. After no more regressors could be removed, the new populations were calculated for these coefficients as well as for an IV regression where the percentage of the population commuting was used as the only regressor. Both approaches led to extreme population values. Between -150.000 and 3.5 million for the single regressor and between -330 million and 4,25 billion for the interactions. Because of these extreme values the percentage approach was discarded in favour of the absolute value approach. Carrying out an OLS regression of the exogenous jobs on the total number of outgoing commuters showed the highest R2 for a buffer of 20 km's. The R2 is only around 10 % however so the instrument is quite weak. Like with the percentage R2, missing data could be a possible cause. For the absolute values three specifications were tried, one with only the single regressor, one with only some demographic and economic interactions and one with more demographic and economic interactions as well as local service and amenity

interactions. For this last specification, the same approach was taken as with the percentage approach in terms of removing statistically insignificant regressors. The coefficients found from these regressions can be found below in table 2.

VARIABLES	(1) Single regressor	(2) Few interactions	(3) Many interactions
outgoing_commuters	3.452***	-7.403*	9.954**
	(0.550)	(4.240)	(4.730)
1.sted#outgoing_commuters		0	0
		(0)	(0)
2.sted#outgoing_commuters		-0.219	0.424**
		(0.324)	(0.184)
3.sted#outgoing_commuters		-0.641	0.464
		(0.610)	(0.328)
4.sted#outgoing_commuters		-1.121	0.425
		(0.874)	(0.366)
5.sted#outgoing_commuters		-1.416	0.265
		(1.082)	(0.459)
outgoing_commuters#p_15_24_jr			0.0315
			(0.0563)
outgoing_commuters#p_25_44_jr			-0.0747
			(0.0626)
outgoing_commuters#p_45_64_jr			-0.0912*
			(0.0482)
outgoing_commuters#p_65_eo_jr			-0.0724
			(0.0489)
outgoing_commuters#woz		-0.00157	0.00311
		(0.00202)	(0.00238)
outgoing_commuters#p_huurwon		0.138***	-0.0153
		(0.0362)	(0.0199)
outgoing_commuters#p_arb_pp			-0.0333
			(0.0349)
outgoing_commuters#m_hh_ver			-0.0111**
			(0.00480)
outgoing_commuters#av10ziek_i			-0.0848**
			(0.0384)
outgoing_commuters#av5_daglmd			0.00503***
			(0.00181)
outgoing_commuters#av5_restau			0.00281***
			(0.00108)
outgoing_commuters#av5_bso			-0.00850
			(0.00631)
outgoing_commuters#av10museum			-0.0919***
			(0.0289)
outgoing_commuters#av10podium			0.0449*
			(0.0266)
outgoing_commuters#av10ondvrt			0.133***
			(0.0419)
outgoing_commuters#av10attrac			0.0755
			(0.0463)
outgoing_commuters#gem_hh_gr		1.911	
		(1.201)	
Constant	3,614	30,814***	17,214***
	(6,054)	(10,583)	(5,571)
Observations	355	355	355
R-squared	0.854	0.905	0.972

Robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Table 2 – alternate specifications regression results

The estimation with few interactions suffers quite a bit from statistical significance issues, the estimations with a single regressor and with many interactions are both mostly significant. Before continuing to the rank sized rule tests. It is valuable to take a look at the interaction coefficients in the many interactions case. More information on the used variables can be found in (CBS, 2021). Interestingly it seems that the coefficients describing the urbanisation grade show that urbanisation grades ranging between weak and strong urbanisation lead to more population per commuter than very strong urbanisation or no urbanisation. This suggests that commuters with a larger group of dependents commute from small to medium large cities towards major cities. The interactions confirm the logical intuition that a large concentration of youth leads to a large population gain per commuter. This large concentration of youth is indicative for a large concentration of families who have a high appreciation for living space and will therefore settle in commuting towns. House value, percentages rental housing and labour participation are quite insignificant but show a vague picture of richer places with a low proportion of working population generating more population per commuter than cheaper places with more rental housing. The coefficient for average household assets is more significant and contradicts this picture. Services and amenities show a very mixed picture. High schools, podiums, attractions, restaurants and shops providing daily foodstuffs all have a positive impact on the amount of population gained per commuter. Hospitals, daycare and museums have a negative impact on the amount of population gained per commuter. These differences are difficult to intuitively explain. It lies outside of the scope of this paper to address these in more depth. So, for now this issue will be left with the unsatisfactory explanation that apparently some amenities and services sort for a public with larger families than other amenities and services. When computing the gained population per commuter per municipality, the coefficients for the many interactions are quite large, ranging between 5 and 12. Interestingly, the largest municipalities have the highest coefficient with Amsterdam being a clear frontrunner. Seeing as the single regressor is quite a bit smaller, it seems like these coefficients are

overestimations. The problem with weak instruments is that the resulting coefficients will show a bias towards the OLS estimator. Seeing as the number of commuters is a very endogenous variable, this bias can turn out quite large. In this case the OLS estimator returns very high coefficients. Because larger cities receive the most commuters however, this overestimation of the population per commuter somewhat balances out to the point of resulting in a rank size coefficient close to one. If this coefficient is reliable however, the added population through labour market effects (extra demand and agglomeration benefits) is shown to heavily outweigh the added population through dependents.

As with the percentage approach both the single regressor and heavy interaction approaches resulted in negative population values for different municipalities. In the case of the single regressor this only happened with one municipality and the negative population was limited to –1154. Since this only applied to one municipality and the population is only slightly negative this could be interpreted as a municipality disappearing; therefore, the observation was deleted. For the heavy interactions, a lot more municipalities ended up with negative populations. Not to such an extreme degree as with the percentage approach but still quite extensive, the lowest negative value was –100.000 for Almere. The upper end estimations were more in line with expectations than with the single regressor percentage approach however, with the population of Amsterdam approaching 2 million and the population of Rotterdam and Utrecht approaching around 1.2 million. Because deleting all negative observations would increase the sum of the population (total population) quite drastically the sum of the negative values was divided proportionally over the municipalities with positive values. After deleting the municipalities for which the population is now set to zero, 258 municipalities remain.

6 – results and discussion

Three separate model specifications were tried. One with the number of outgoing commuters as a single regressor, one with limited added interactions and one with many interactions. Table 2 shows the rank sized coefficients calculated from the new populations estimated with these specifications. The provisional results from the different specifications are conflicting. The estimation with limited interactions shows an urban network that is more evenly distributed than would be expected from Zipf's law, while the estimations with the single regressor and many interactions show a distribution that is more uneven than would be expected from Zipf's law.

VARIABLES	(1) Single regressor	(2) Light interactions	(3) Heavy interactions
lnrank	-1.183***	-0.698***	-1.579***
	(0.0377)	(0.0208)	(0.0811)
Constant	15.68***	13.89***	16.99***
	(0.181)	(0.0912)	(0.367)
Observations	354	355	258
R-squared	0.867	0.802	0.787

Robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Table 3 – Rank size coefficients commuter model

The R^2 values of the different estimations are not close to 1 however indicating a lack of a power law distribution. Figures 8 to 14 show that the lower tails of the distributions deviate from the power law while the upper tails are power law distributed. The values emerging when the deviating lower tails are taken are shown in table 3. The new values for the single regressor and many interactions estimations now approach the strict Zipf's coefficient of –1. The estimation from the limited interactions is smaller than that of the original municipalities. As can be seen in table 2, the statistical significance is quite an issue with the limited interactions. Therefore, it is likely that the single regressor or the many interactions estimations give a better estimate of the true power law coefficient for the Netherlands. Table 3 includes two different estimates of the many interactions specification. Figures 14 and 15 show

how deleting parts of the lower tail results in a distribution that resembles a power law but is a bit biased. Both the largest and the smallest cities are smaller than would be expected while the middle of the distribution is slightly larger than would be expected. The deviations from the power law are slight but this might suggest that the Dutch urban system is not ideally described by a power law but with a similar nonlinear distribution. While the eventual results of the single regressor and many interaction estimations are similar, there is a notable difference. While the power law coefficient can be estimated from 250 municipalities for the single regressor estimation, the many interactions estimation for 150 municipalities is still biased. At this point there are less than half of the original starting municipalities remaining. The first 97 municipalities were removed because the model left them with negative populations. It is unrealistic that 97 municipalities would actually disappear if the network of dependencies would be untangled. (Krugman, 1991) models his regions as having a mobile (industrial) and an immobile (agricultural) population. It is likely that an agricultural population along with a village offering basic services would remain. Since these municipalities would be situated at the very tail end of the distribution, the power law would not hold for them, and they would therefore not be considered.

	(1)	(2)	(3)	(4)
VARIABLES	Single regressor	Light interactions	Heavy interactions top 200	Heavy interactions Top 150
lnrank	-0.955***	-0.592***	-1.204***	-1.087***
	(0.0146)	(0.00837)	(0.0387)	(0.0320)
Constant	14.81***	13.47***	15.66***	15.28***
	(0.0694)	(0.0410)	(0.174)	(0.135)
Observations	250	300	200	150
R-squared	0.987	0.980	0.956	0.971

Robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Table 4 – corrected rank size rule coefficients

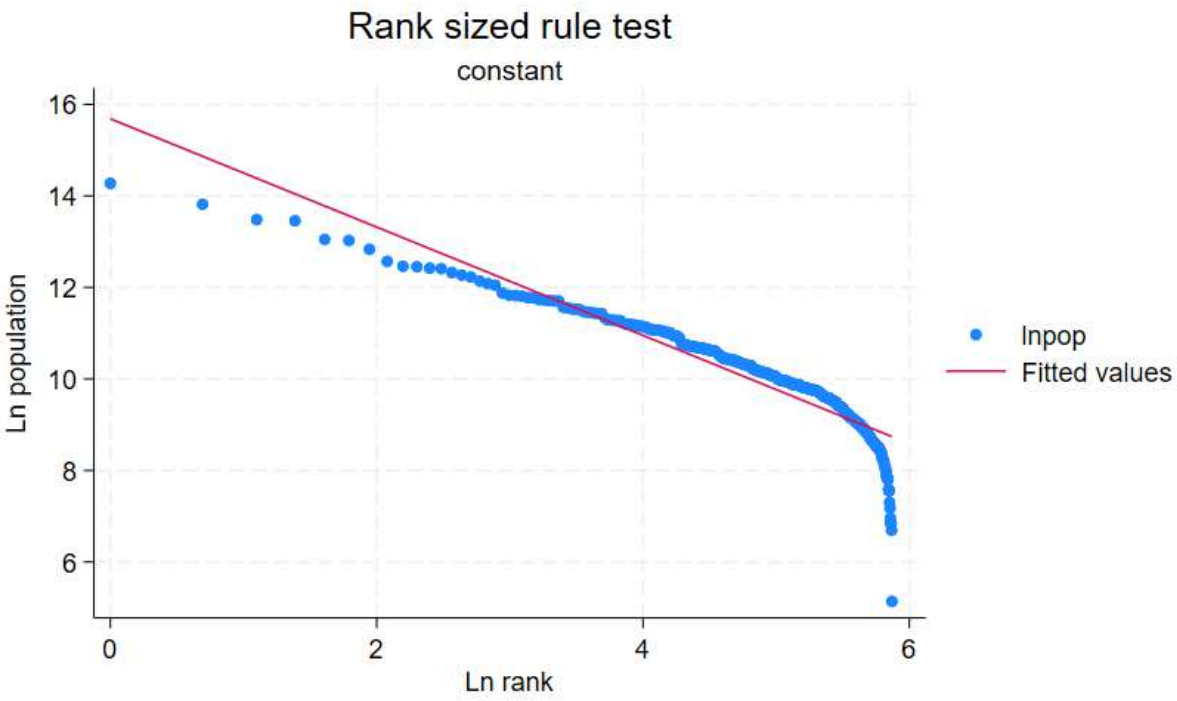

Figure 9 – rank size rule test from the single regressor commuter model – slope –1.183

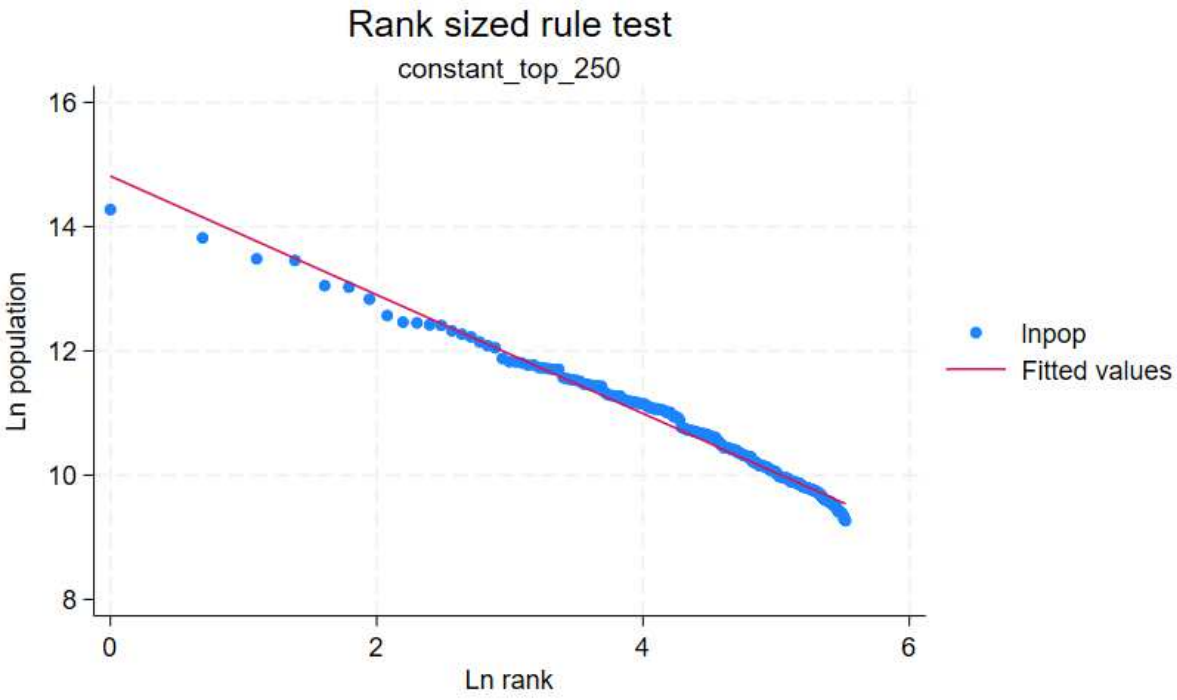

Figure 10 – rank size rule test from the single regressor commuter model top 300 – slope –0.955

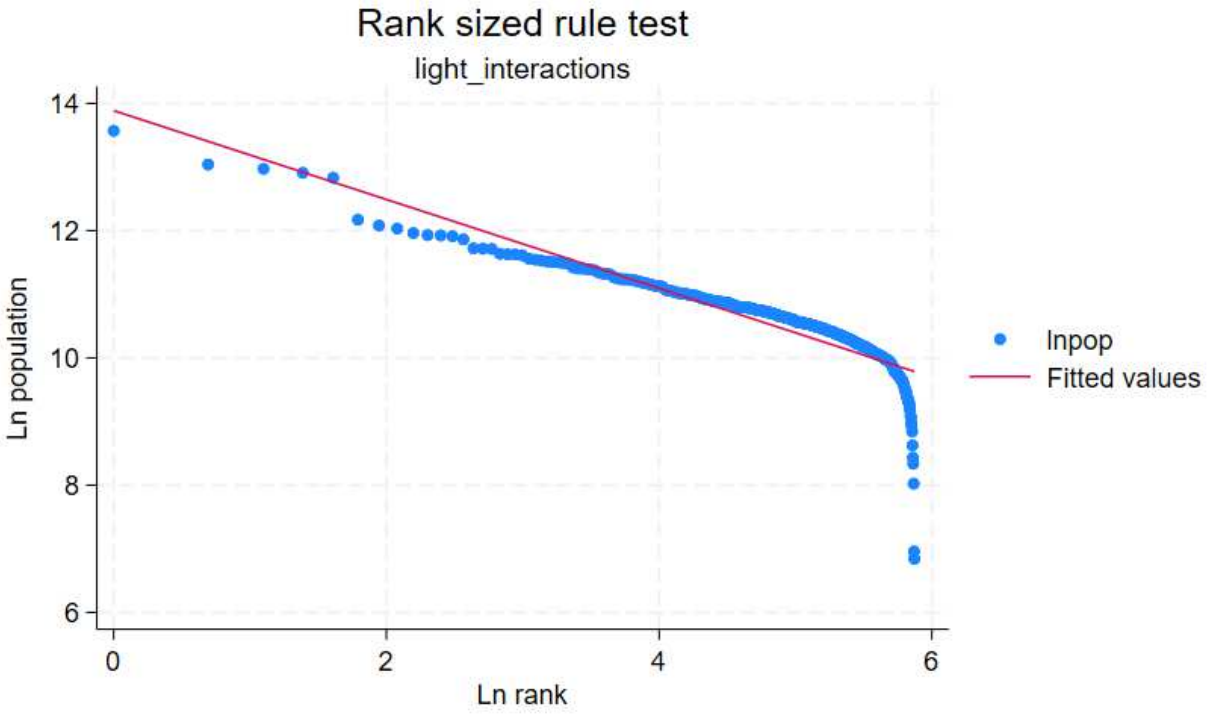

Figure 11 – Rank size rule test from the few interactions commuter model – slope –0.698

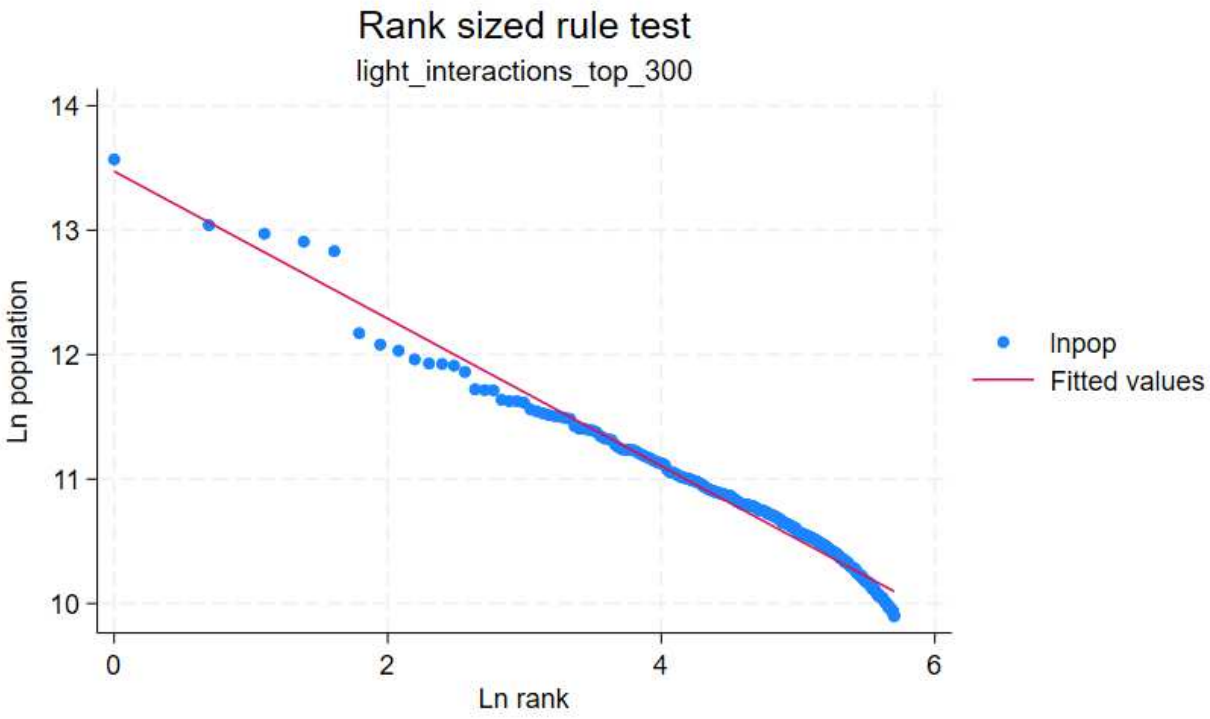

Figure 12 – rank size rule test from the many interactions commuter model – slope –0.592

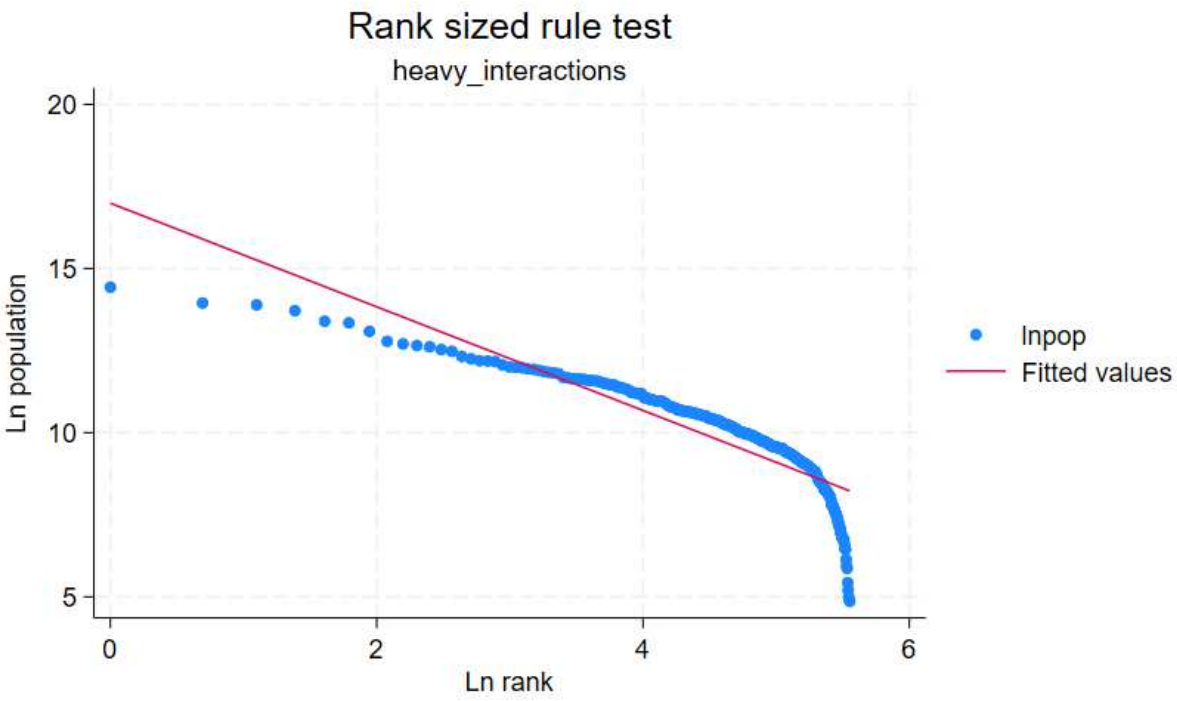

Figure 13 – rank size rule test for the many interactions commuter model – slope –1.579

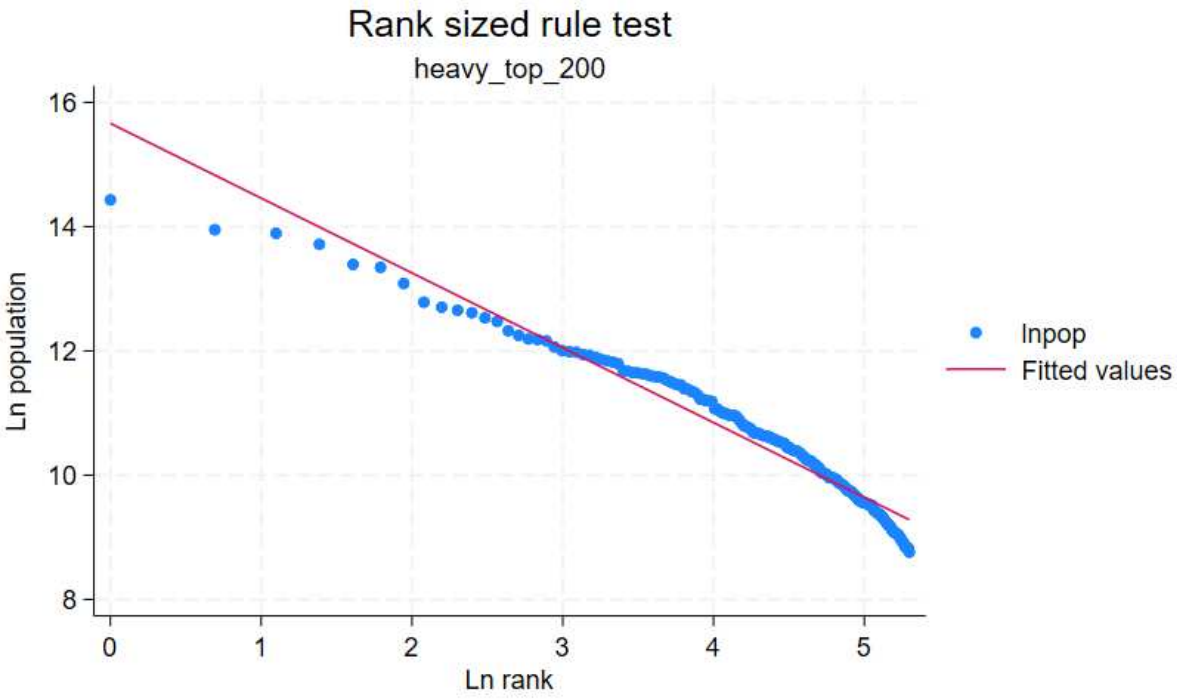

Figure 14 – rank size rule test for the many interactions commuter model top 200 – slope –1.204

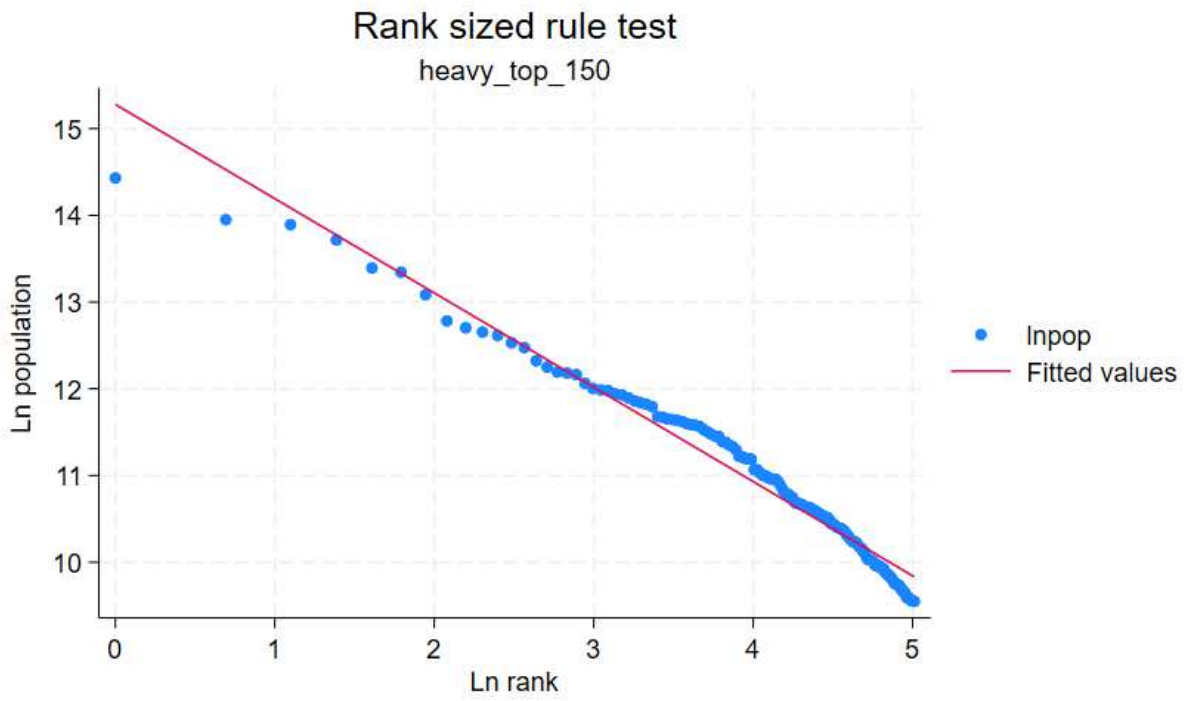

Figure 15 – rank size test for the many interactions commuter model top 150 – slope –1.087

Because of the weak instruments, the internal validity is somewhat compromised. It is therefore not possible to pinpoint an exact rank sized coefficient for the Dutch urban system. The two most valid estimations carried out however lie in the centre of the range (Nitsch, 2005) describes. It is thus possible to conclude with a high degree of certainty that the interconnectedness of the Dutch urban network and the commuting patterns to be precise are the cause for the Netherlands deviating from the rank sized rule and that the populations corrected for commuters and the population following them follows Zipf's law.

Compared to the rank sized coefficient estimated from the jobs, the commuter model shows a more unequal urban distribution. This holds for both the single regressor as the many interactions estimation. Because the working population and dependents are directly correlated to jobs, the difference between the jobs estimate and the single regressor estimate suggests that the non-working independent population is heavily concentrated in larger cities.

This paper sparks some questions and research suggestions. The most pressing of these is whether this research outcome is a fluke. If this methodology were used to check the urban distribution of other heavily interconnected countries, would the rank sized rule follow? There are also some potential improvements to the model. The current model constructs new population on the basis of their existing jobs. It is however doubtful that everyone would choose to remain with their current job if they had to live where they worked. A model that redistributes population based on both the producer and consumer side would be an improvement on the model. The main issues this model has with internal validity stem from the somewhat weak instruments. Using more modern data on jobs would likely strengthen the instrument and the internal validity. This model was run on December 2018 job and commuter data to avoid distortions caused by the covid-19 pandemic and its medium-term effects on commuting and place of living choice. The 2018 data had some problems with missing values however. The modern data on the other hand is more complete. This paper is also relevant for policy makers. This paper provides a method to scale interdependencies between municipalities. This is very relevant in questions surrounding local taxation. In the current system, smaller municipalities go unpunished for freeriding on services and amenities provided through taxes levied on core city inhabitants. The scaled interdependencies allow for the making of policy concerning fairly sharing tax burdens.

7 - conclusion

At the start of the paper, the question was asked whether the very equal distribution of population in the Netherlands estimated from municipality populations was an accurate representation of the urban system in the Netherlands. A reformulation of that question is: 'what city definition most accurately describes the urban structure in the Netherlands?'. The analysis of Zipf's law estimations through several alternative city definitions has shown that estimations utilizing definitions more catered to estimating urban networks show a more unequal distribution. These metropolitan definitions are limited however in that they remain tied to geography. Utilizing a system that discards these spatial requirements allows for a more thorough deconstruction of the Dutch urban system. To this end the commuter model was created. A model that redistributes the population of Dutch municipalities along commuter flows. The model furthermore accounts for local population the commuters generate at their place of living. These consist of direct dependents, non-working partners and children, jobs generated through added demand, healthcare workers for instance and jobs generated through agglomeration benefits. Estimating Zipf's law through the new populations generated by the commuter model results in coefficients of about –1. Keeping in mind that the internal validity of the estimation is contentious due to weak instruments, this result allows the cautious conclusion that the rank size distribution of urban networks in the Netherlands does in fact correspond to Zipf's law. The rank size coefficient has also been estimated from the number of jobs per municipality. While showing a distribution that is more unequal than that of population from municipalities, it is still quite some distance from the estimation through the commuter model. This suggests that outside of commuters and direct dependents, labour market effects and independent non-working population play a large role in determining the urban distribution in the Netherlands.

If Zipf's law holds for the Netherlands, (Gabaix, 1999) suggests that Gibrat's law would also hold. The implications of Gibrat's law holding are quite significant. Gibrat's law states that the mean growth

and standard deviation of growth are independent of size. This means that past success is no predictor in future success. Essentially, growth is random in the long term. The Netherlands is famous for attempting to plan and control land use and urban development to a degree scarcely found anywhere else in the world. If Gibrat's law holds, this idea of control vanishes. Gibrat's law holding forces us to reconsider certain decisions. If regional growth is uncontrollable in the long term, what use is investing in shrinking regions for instance? Does this mean the Netherlands should quit their heavily planned and ordered approach to spatial planning, since a future of random growth is unpredictable? I do not think so, growth may be random in the long run, but to quote the great John Maynard Keynes: "In the long run we are all dead".

References

Brakman, S., Garretsen, H., & Van Marrewijk, C. (2020). *An introduction to geographical and urban economics*

A spiky world (3 ed.). Camebridge University Press.

Brush, J. E. (1966). Walter Christaller. Central Places in Southern Germany. Translated by Carlisle W. Baskin. Pp. 230. Englewood Cliffs, N.J.: Prentice-Hall, 1966. $9.95. *The ANNALS of the American Academy of Political and Social Science*, *368*(1), 187-187. https://doi.org/10.1177/000271626636800132

CBS. (2021). *Toelichting Wijk- en Buurtkaart 2021*. CBS. https://www.cbs.nl/nl-nl/longread/diversen/2021/toelichting-wijk-en-buurtkaart-2019-2020-en-2021?onepage=true

CBS. (2022a). *Banen van werknemers naar woon- en werkregio (2014-2020)*. https://opendata.cbs.nl/portal.html?_la=nl&_catalog=CBS&tableId=83628NED&_theme=781

CBS. (2022b). *De toelichting op de SBI 2008 - versie 2018 Update 2022*. Retrieved from https://www.cbs.nl/nl-nl/onze-diensten/methoden/classificaties/activiteiten/standaard-bedrijfsindeling--sbi--/de-toelichting-op-de-sbi-2008-versie-2018-update-2022

CBS. (2023). *Banen van werknemers in december; economische activiteit (SBI2008), regio*. https://opendata.cbs.nl/statline/portal.html?_la=nl&_catalog=CBS&tableId=83582NED&_theme=246

CBS, & Kadaster. (2021). *Wijk- en buurtkaart 2019*. https://www.cbs.nl/nl-nl/dossier/nederland-regionaal/geografische-data/wijk-en-buurtkaart-2019

De Groot, H., Marlet, G., Teulings, C., & Vermeulen, W. (2010). *Stad en Land*. https://www.cpb.nl/sites/default/files/publicaties/download/bijz89.pdf

Gabaix, X. (1999). Zipf's Law for Cities: An Explanation. *The Quarterly Journal of Economics*, *114*(3), 739-767. https://doi.org/10.1162/003355399556133

Gabaix, X., & Ioannides, Y. M. (2004). Chapter 53 The evolution of city size distributions. In (pp. 2341-2378). Elsevier. https://doi.org/10.1016/s1574-0080(04)80010-5

Glaeser, E. L., Kolko, J., & Saiz, A. (2001). Consumer city. *Journal of Economic Geography*, *1*(1), 27-50. https://doi.org/10.1093/jeg/1.1.27

Groot, S., De Groot, H. L. F., & Veneri, P. (2012). The Educational Bias in CommutingPatterns: Micro-Evidence for the Netherlands. https://tinbergen.nl/discussion-paper/3755/12-080-3-the-educational-bias-in-commuting-patterns-micro-evidence-for-the-netherlands

Jiang, B., & Jia, T. (2011). Zipf's law for all the natural cities in the United States: a geospatial perspective. *International Journal of Geographical Information Science*, *25*(8), 1269-1281. https://doi.org/10.1080/13658816.2010.510801

Krugman, P. (1991). Increasing Returns and Economic Geography. *Journal of Political Economy*, *99*(3), 483-499. https://doi.org/10.1086/261763

Liu, H., & Liu, W. (2009). Rank-Size Construction of the Central Place Theory by Fractal Method and Its Application to the Yangtze River Delta in China. *Proceedings - International Conference on Management and Service Science, MASS 2009*. https://doi.org/10.1109/ICMSS.2009.5301777

Marlet, G., & Van Woerkens, C. (2014). De nieuwe gemeentekaart. *Atlas voor Gemeenten*.

Nitsch, V. (2005). Zipf zipped. *Journal of Urban Economics*, *57*(1), 86-100. https://doi.org/10.1016/j.jue.2004.09.002

Simon, H. A. (1955). ON A CLASS OF SKEW DISTRIBUTION FUNCTIONS. *Biometrika*, *42*(3-4), 425-440. https://doi.org/10.1093/biomet/42.3-4.425

Steindl, J. (1965). *Random processes and the growth of firms*. Ashgate Publishing.

Appendix 1 - Lotka vs Pareto specification

This paper makes use of the Lotka specification because it is more intuitive to interpret and because earlier literature on Zipf's law in the Netherlands uses the Lotka specification. Most literature prefers the use of the Pareto specification however. Volker Nitsch equalizes these estimates to one metric, the Pareto specification. To do so, he calculates the inverse from the Lotka specification estimates. This is not completely econometrically sound however seeing as OLS minimizes vertical errors and taking the inverse converts vertical errors to horizontal errors. In the case of small standard errors usually found in rank size rule estimation, this shortcut will lead to only small deviations. A short test reveals that the inverse of the Lotka estimation for the top 200 municipalities in the single regressor commuter model results in an estimation 0.01 point higher than the Pareto estimation. From this result the two specifications can reasonably be compared by using the rule of thumb that the estimations are inverses of each other in the knowledge that this is ultimately not completely true.