A Data Science study of Small Rural Towns in the Netherlands

IBM Data Science Certificate - Capstone Project

Laura Bongaardt
Delft, Netherlands - August 23, 2021*

I. Introduction

It is often said that rural towns in modern society are struggling. Usually it concerns historical places that once relied on manual labour jobs or agriculture for income. With automation and globalisation many of those jobs are gone and while cities found new roles with trade and innovation, rural towns are still lacking behind. There are many small rural towns in the Netherlands as well, but we are a small country and even towns that we call very remote are still relatively close to bigger places. In this text we want to take look at the status of Small Rural Towns (SRTs) in the Netherlands and what the impact is of these trends.

This will be a study from a data science perspective. By using publicly available data from the Central Bureau of Statistics (CBS), Metatopos.eu and the Foursquare API, we can compare key metrics of small rural towns with cities and with other big towns. We can examine which towns are succesful and will try to identify factors that can help other communities.

This text is intended for (local) governments to create policies that stimulate local economies, for business men that want to invest in rural towns, people that are looking to move to the countryside or for residents that want to help their communities.

The objectives are:

1) Identify and locate Small Rural Towns of interest, based on Population Size and Distance to a city.
2) Make an Assessment of Small Rural Towns, relative to cities and Big towns, based on Population Growth and Average Income.
3) Score Small Rural Towns indidually, based on Population Growth and Average Income.
4) Create Neighborhood profiles for Small Rural Towns
5) Examine relationships between performance and Neighborhood features

*Pictured:Town of Lekkum, credit@TimoJepkema

II. Data

i. Central Bureau of Statistics (CBS)

- *Table 84992NED*: List of registered places, cities and towns, in each municipality in the Netherlands. Dated Jan 1, 2021.
- *Table 84799NED*: Population size and area size for each 4-digit postalcode in the Netherlands. Data collected for year 2020.
- *Table 83502NED*: Population size for the period 1998-2020, for each 4-digit postalcode. Last update Sep 4, 2020.
- *Table 84286NED*: Average income for each 4-digit postalcode. Dated year 2018.

ii. Metatopos.eu

List of all 4-digit postalcodes belonging with each registered place, city or town. Dated 2021.

iii. Foursquare API

List facilities and venues in the neighborhood of selected small rural towns.

III. Methodology

i. Gathering Data

We start by collecting data from various sources. The Central Bureau of Statistics (CBS) collects many different indicators and statistics and makes them publicly available through their data portal: *https://opendata.cbs.nl/*. Data can either be downloaded in csv format or requested with an API. Through Table 84992NED we obtain a list of all 2500 registered places the Netherlands, along with their municipality, province and place code. Latitude and Longitude for each place are obtained using the GeoPy package.

Table 84799NED contains the population and area size, listed for each neighborhood-block ('wijk') along with the most dominant 4-digit postalcode in that block. By scraping from *Metatopos.eu*, which obtains its data from the postal-service, we get a list of all postalcodes belonging with each place. The designations of registered places, postalcode-areas and neighborhood blocks are done by different organisations and they can overlap. However, as a general rule, each block or 4-digit postalcode is to be contained within a registered place in its entirety. This means that we can group the postalcodes for each place and summate the metrics, to get a reasonable estimate of the population and area size of each place.

The result is a dataframe which contains registred places, with municipality and province, their respective population and area size, their longitude and their latitude. Distances and population density can be calculated from these as well.

ii. Selecting Small Rural Towns

The next step is to select places of interest from the dataframe. We are looking for small towns located at some distance away from a city center. What constitutes 'small' or 'some distance away' has to be decided on the available data.

In Figure 1, population size and density are plotted for each place, ordered from small to large. We can see that there are a few places that have a lot of people. These are the big cities such as Amsterdam, the capital, with over 800.000 people. However the vast majority are much smaller places and the medium size is actually just over 1500 people.

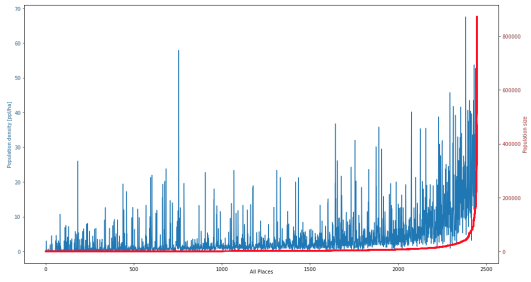

Figure 1: *Population Size and Density per Place*

We divide the set into 3 categories:
1) Cities, with over 104.000 people
2) Big towns, between 3000 and 104.000
3) Small towns, between 300 and 3000 people

Places with fewer then 300 people are discarded. These often involve isolated houses or very scarcely populated areas and they are fluctuating too much to get a good measure.

We now look at the location of the small towns. Using the GeoPy package, we can calculate their distance to the nearest big town and city. Shown in figure 2 is a density plot of the distances, in km, against the proportion of the samples. The Netherlands is a small country and many of the places are located in the close proximity of larger towns. We want to select rural towns, but still get a decent sample size. As a threshold, places were selected with a minimum distance of 15km to the nearest city and a minimum of 4km to the nearest big town.

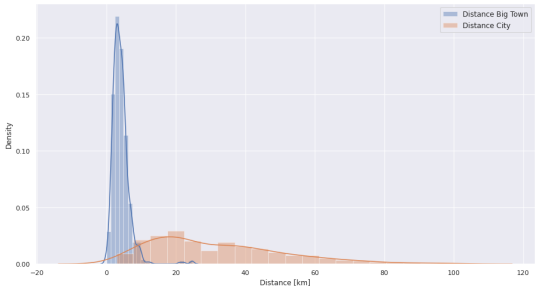

Figure 2: *Distances Small towns to nearest City and Big town*

Using the afore mentioned criteria, we find 249 Small Rural Towns (SRTs). These are places with a population between 300 and 3000 people, located at a minimum 15km from the nearest city and 4km from the nearest big town.

In figure 3, these towns are plotted on a map of the Netherlands. We can see SRTs are spread out, but most are in the North, East and South parts of the country.

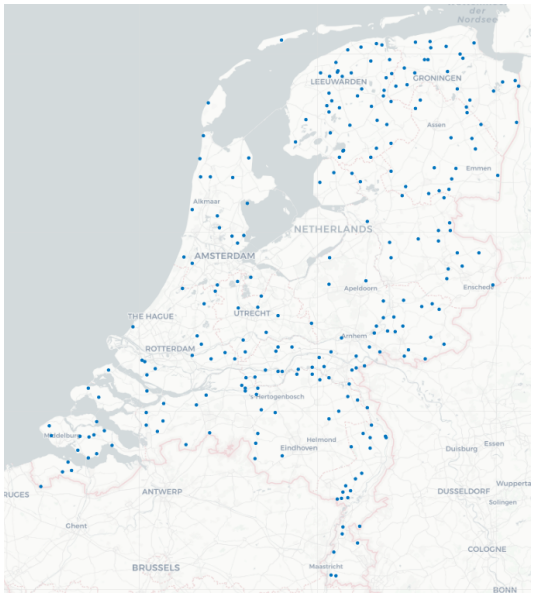

Figure 3: *Selected Small Rural Towns*

iii. Assesment of Small Rural Towns

After selecting the SRTs, we want to get an assessment of their status. Population growth and income levels are two basic indicators which can be obtained from tables from the CBS. Population numbers are listed in table 83502NED for the period between 1998-2020. Average Income for 2018 was found in table 84286NED. 2018 is slightly out of date, but since 2020 or 2021 are not entirely representative either, due to the COVID-19 pandemic, we will use this as a comparative indicator.

Data from these tables is listed per neighborhood block along with the most dominant 4-digit postalcode in that block. We can link postalcodes to the corresponding place and then group them together to get the indicators for each place. In case of Average Income, we take the average of the values, which will get us the average for the entire place. To get a general assessment of the performance of SRTs, we will compare these indicators with cities. We could not find data for all SRTs, so we had to continue with 244 places.

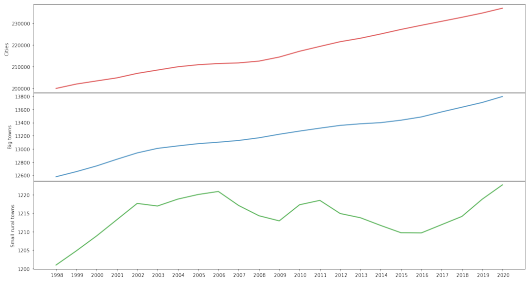

Figure 4: *Mean Population 1998-2020 per Category*

In figure 4 is the average population plotted, for the 3 types of size categories, over the last 22 years. We can see clear differences between them. Cities grew from 200.000 to 230.000, or 15%. Big towns grew from 126.000 to 138.00, or 9.5%. Small rural towns however, on average, grew only by about 20 people over the last 22 years. This constitutes just 1.7% and the line only moves due to fluctuations in the data.

When we look at the Average Income for 2018, as shown in figure 5, the differences are smaller. In 2018, the national average was €26.600. The medians of cities and big towns are around that figure as well. For SRTs, the median is slightly lower, just over €25.000. We do see a greater spread among big towns and SRTs. There are outliers, but also places that are €6.000 below the national average.

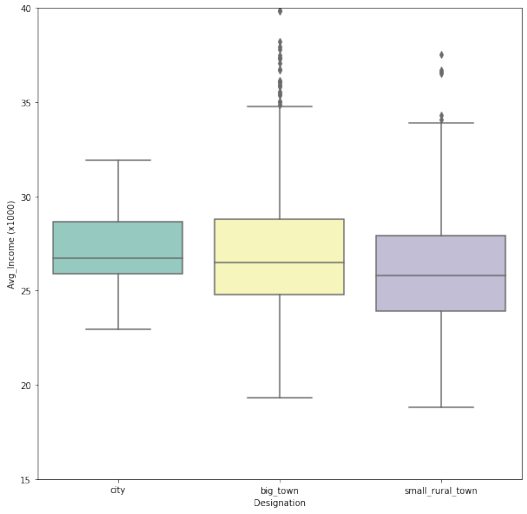

Figure 5: *Average Income 2018 per Category*

iv. Performance scores SRTs

In the previous section we looked at average figures, but there is a significant spread and there are large differences between the towns.

In figure 6, Population Growth (PG), in percentages, is plotted against the Relative Income (RI) for SRTs. RI is the average income subtracted with the national average of €26.600. This gives us a better intuition of the welfare of a town.

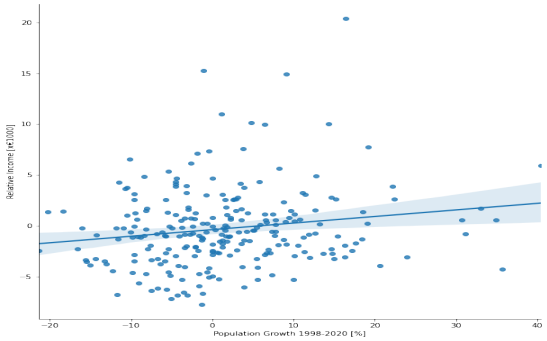

Figure 6: *Population Growth vs Relative Income*

We can see there is a wide range in PG and RI, but we could find no meaningful relationship between them. The correlation coefficient was found to be 0.17. To assess the performance of SRTs, we normalise PG and RI and use a weighted scoring. PG is deemed more important, so this gets a weight of 2. RI gets a weight of 1. Multiplying by 10, we get a resulting score on a scale from 0 to 10. Peformance score for each SRT is plotted in figure 7.

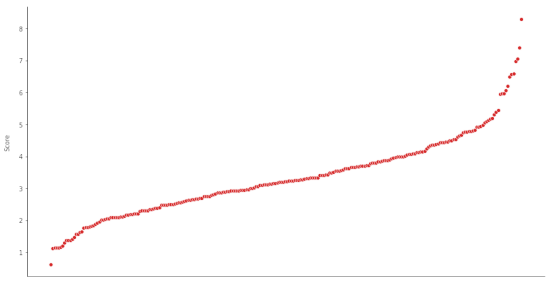

Figure 7: *SRT Performance Scores*

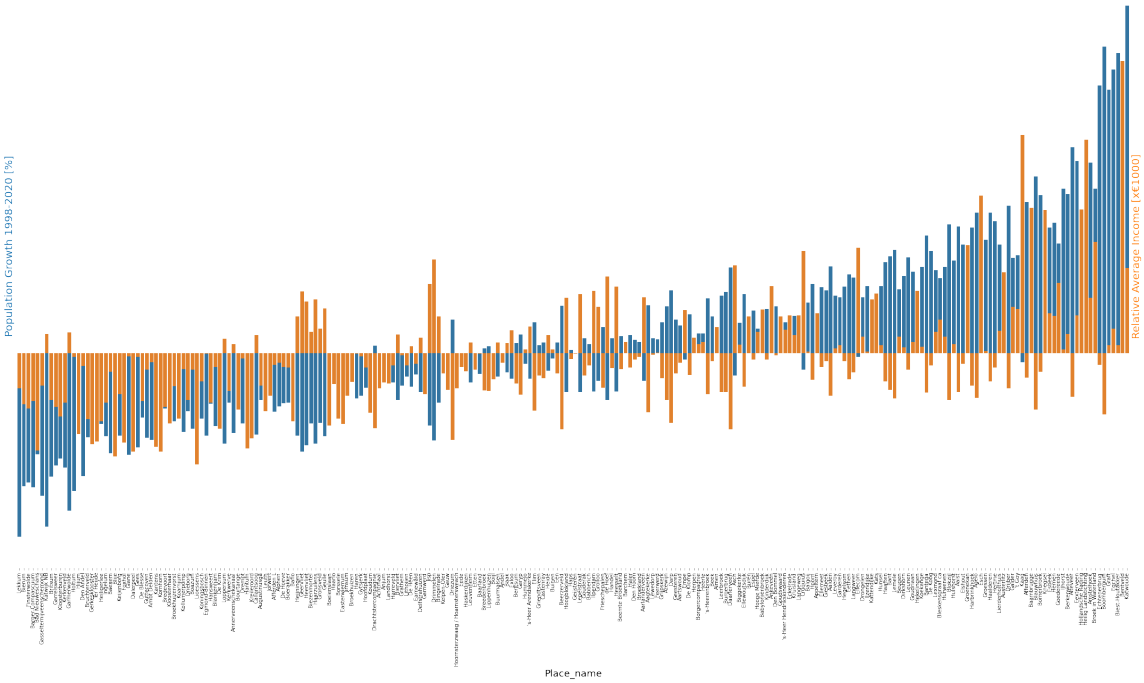

Figure 8: *Growth and Income for Small Rural Towns*

In figure 8, we see a barplot of all SRTs, with their respective Population Growth and Relative income, ordered by Performance score.

On the far left of the spectrum is the town of Lekkum, with a Score of 0.6. Its population shrank from 515 to 405, a decline of 21%, while its Average Income in 2018 was €2451 below the national average.

On the other end we have Katwoude, with Score 8.3. Its population grew from 235 to 330, an increase of 40%. Average Income was €5948 above national average.

With these scores we can categorize the SRTs into 3 approximately equal groups. In figure 9, they are plotted on a map with score labels. Green are places with a score above 4.2 and that are doing well. Red are SRTs that do badly, with a score below 2.7. And Yellow are the places in between.

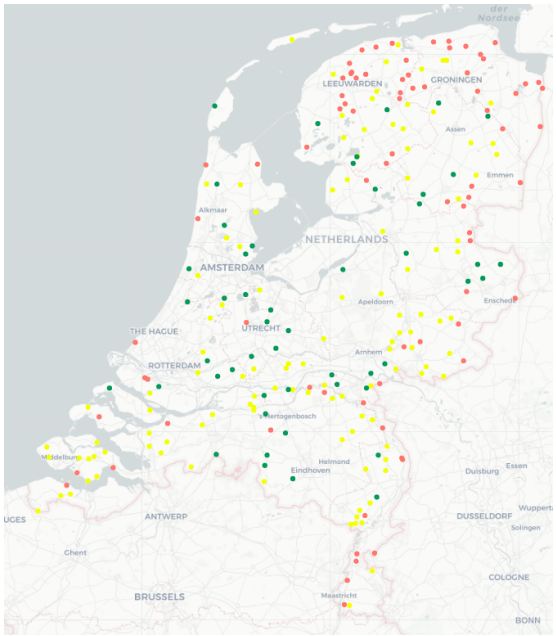

Figure 9: *SRTs with score labels*

v. Profiling SRT Neighborhoods

From the map in figure 9, we can see that there are more places in the North that do badly, while there are more places in the West that are doing well. The West is also the region where most of the cities, big towns and provinces are that are generally doing well. So we can logically expect a correlation between the performance of a province and the performance of the SRTs within it. But the location of a SRT is not something we can control and we are more interested in potential factors that we might be able to influence, such as access to facilities or stimulating certain industries. In this section we will use the FourSquare API to build a neighborhood profile for each SRT. With these profiles, we will then try to find factors that have a relationship with Performance Score.

Access to facilities could be a factor that can influence the succes of a town. If it is, SRTs that are closer to facilities might have an higher Performance Score. To test this, we will list the distance to a number of facilities for all 244 SRTs: Hospital, Train station, Elementary school, High school, Police and Fire station. We will use FourSquare API to find the facilties and the GeoPy library to find the nearest in the Area. FourSquare API could not find them all, so we have added a few by hand.

A summary of the data is shown in figure 10. We see boxplots of the distances to facilities, in meters, for 244 places. We can see that 4 of the categories are in a relatively narrow range.

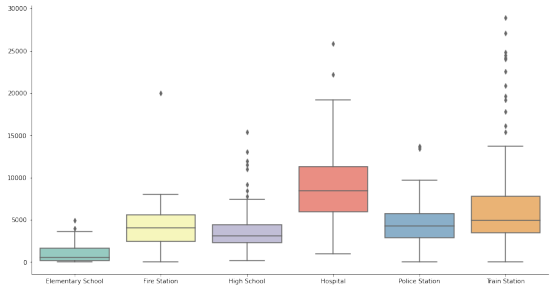

Figure 10: *Distance to nearest Facility*

Distance to Hospital and Train station have the most variation.

The second factor that we would like to examine is whether the presence of certain industries affects the performance of a town. To test this, we use FourSquare API to list all venues, in 9 different categories, within a radius of 1500meter of the center of the SRT. In total we found 5583 venues. In figure 11, we see the range of the number of hits per category.

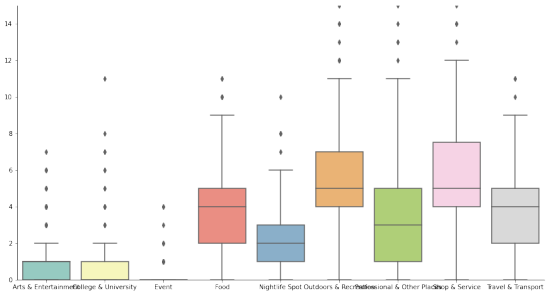

Figure 11: *Number of Venues per Category*

We can see we found very few Event venues, so we drop that category. The category 'Professional & Other Places' is interesting to look at. In Figure 12, we see a sunburst chart of this category.

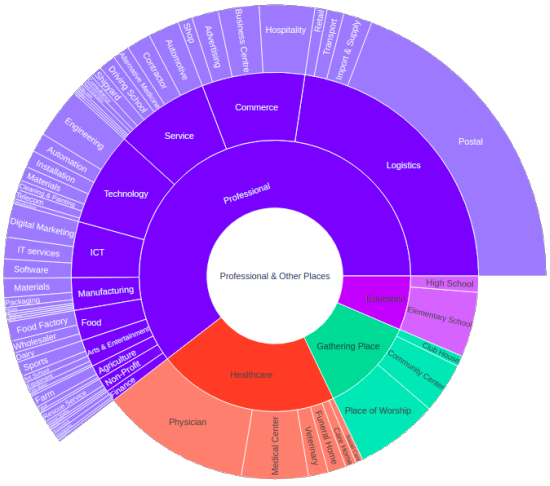

Figure 12: *Professional Places within a 1500m radius*

vi. Feature Analysis

We now try to find possible relations between the features, distance to facilities and venues in the area, and the Performance Score.

We will calculate correlation coefficients between each of the features and the Performance Score. If a coefficent is close to 1 or -1, this would indicate a strong relationship.

IV. Results

In figure 13, we see a correlation map of Pearson's correlation coefficients between distance to facilities and the Performance Score (on the bottom row). Unfortunately, we can not find any meaningful relationship. The strongest coefficient is -0.16. Possible cause might be that there is not enough variation between the SRTs, for it to be a factor. Another possibility is that none of the facilities are far enough for it to matter. The Netherlands is a small country and even 30km, the furthest distance to a hospital, is only a 30 minute drive.

Next, we look at correlations between venues in the neighborhood and Performance Scores, figure 14. We can see a moderate relationship between certain venue types. For example a coefficient of 0.66 between Food and Travel & Transport venues. This suggests that when there are a lot of Travel venues, such as hotels, we can also expect more Food venues, such as Restaurants, in that area.

However we could not find any significant relationship between the presence of venues in an area and Performance Score. The strongest number is only 0.2. One possible reason is that our data was not detailed or extensive enough.

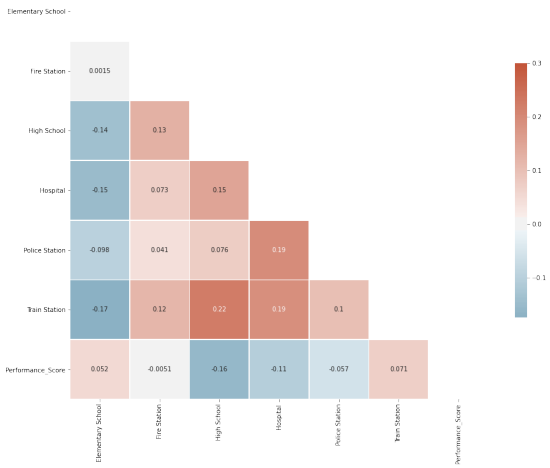

Figure 13: *Correlation map Facility Distances*

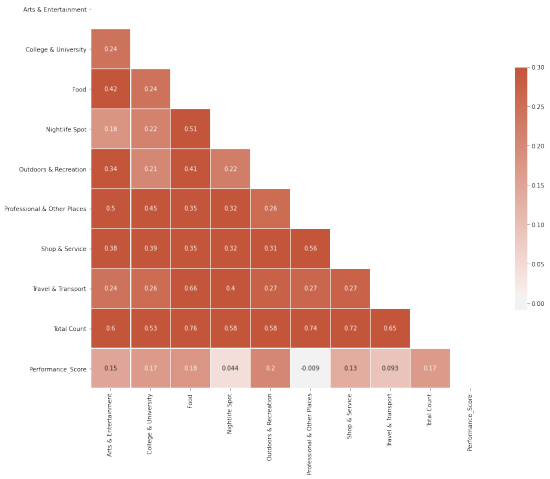

Figure 14: *Correlation map Venues*

V. Discussion

The purpose of this text was to look for relevant factors that might influence the performance of a town. Unfortunately we could find no meaningful relationship. All correlation coefficient were 0.2 or less, indicating a very weak relationship. We did not expect this result.

Possible reasons why distance to facilities is not a factor might be because there is too little spread between the towns and therefore it doesn't have an influence. Another reason might be that the absolute distances are still all quite close, because the Netherlands is a small country. 30km, the largest distance to a hospital, is still only a 30minute drive.

The presence of certain venue types also didn't had any effect. It might be our data wasn't detailed or extensive enough. Another reason might be that again, because the Netherlands is a small country, venues in the near vicinity are not very important as towns can rely on bigger places for employment. This would also correspond with the absence of a meaningful relationship between Average Income and the Population Growth of a town. It appears that income is not the stongest motivation for people leaving a town.

VI. Conclusions

Using publicly available data and the data science methodology, we could reach the following conclusions:

- We were able to identify 244 Small Rural Towns in the Netherlands.
- From the Assesment, we found that SRTs are falling greatly behind in Population Growth, compared to Cities and Big towns. But we could find no significant difference between Average Incomes.
- We scored the individual towns on a scale from 0 to 10 and discovered a large spread in the figures.
- We've created Neighborhood profiles.
- From the correlation matrices we could not find any meaningful relationship between any of the features and the Performance Score.