FACULTEIT ECONOMIE
EN BEDRIJFSKUNDE

TWEEKERKENSTRAAT 2
B-9000 GENT
Tel. : 32 - (0)9 – 264.34.61
Fax. : 32 - (0)9 – 264.35.92

WORKING PAPER

Combining growth and level data: an estimation of the population of Belgian cities between 1880 and 1970

Authors

Stijn Ronsse Samuel Standaert

January 2017
2017/927

D/2017/7012/01

Combining growth and level data: an estimation of the population of Belgian cities between 1880 and 1970

Stijn Ronsse
Department of Economics
Ghent University
Tweekerkenstraat 2, 9000 Gent, Belgium
Stijn.Ronsse@UGent.be

Samuel Standaert[1]
John E. Walker Department of Economics
Clemson University
237 Sirrine Hall, SC 29634 Clemson, USA
sastand@clemson.edu

Keywords

Demographic growth; State-space model; City growth; Belgium

Abstract

Economists that study long-term changes during the 19th and 20th century are fundamentally restricted by the availability of qualitative data, as the latter is often inversely proportional to quality. This is further compounded by administrative changes that alter what exactly is being measured over time as well as an overall decrease of data availability the further we go back in time. This is particularly inconvenient in historical population data, as census data is often only available ever decade. As a result, researchers are forced to either impute qualitative data, or otherwise combine datasets of varying quality in some way. In this article, we demonstrate the versatility of state-space models in addressing these problems, enabling us to compose large data series of a high quality. Moreover, unlike more simple techniques it also provides an estimate of the reliability of the results, allowing any subsequent analyses to take this into account. We illustrate this by combining growth and level data on the population of Belgian cities into a dataset that contains yearly estimates of the population of over 2600 cities from 1880 to 1970.

Keywords: Population, Data quality, State-space model, Bayesian econometrics

1. Introduction

Research on population and the rate of its growth has always held a central position in economics and economic history. Not only does this data hold information on the demographic structure, it is also informative on economic, political and geographic changes. However, the potential insights that it offers are restricted by the availability of qualitative data. Population censuses are a highly qualitative

[1] Corresponding author

source of data that is frequently used, but are typically only taken every decade. To circumvent this availability problem, population levels are often interpolated, examples of which can be found in Gonzalez-Val (2014). While this approach might work relatively well in years when population growth remains stable (e.g. those without administrative changes or wars), it cannot provide reliable growth rates. Furthermore, without knowing the actual population levels, it is impossible to know for certain that population remained stable during the interpolated years.

This can be a limitation for researchers. For example, most people interested in regional economics are familiar with the empirical regularities that are increasingly studied in the field, i.e. Zipf's and Gibrat's law. These hold that the size of cities follows a power law and that the relation between growth rates and initial population size is stochastic, respectively. While initial studies mostly confirmed the existence of both laws,[2] a recent wave of research disputes those findings. Using historical, un-truncated datasets that includes all geographic entities instead of only the largest cities, they (partially) reject Zipf's and Gibrat's laws (Klein, 2014; Gonzalez-Val, 2014).[3] The rejection of these empirical regularities is important as it indicates that population growth is not stochastic, but needs to be explained by deterministic models. However, identifying the relevant dynamic determinants and identify for example structural breaks is neigh impossible using the sparse census data. In other words, population data needs to be much more detailed –ideally containing yearly observations for all municipalities over long stretches of time– both in order to more fully explore the empirical regularities and to test deterministic models that can explain the observed patterns.

One way of overcoming this problem is to use additional data sources on population. In this article, we demonstrate how this can best be achieved even when these datasets have markedly different characteristics and reliability. By way of case study, we focus on the Belgian population data. Specifically, we show how to combine the 10-yearly population censuses with the yearly growth data coming from sources such as parish or population registers that track births, deaths and migration. The Belgian government collected yearly questionnaires in which each municipality had to report changes to the population register: the *mouvement*. However, when the yearly population data is constructed by adding these yearly changes to the census data of that decade, it quickly becomes clear that both data sources are not entirely consistent. Figure 1 illustrates this by reconstructing the population of the Belgian capital, Brussels, from 1880 to 1970. The black crosses show the census data, and the interpolated census data are indicated by the red dotted lines. The yearly population levels based on the *mouvement* data (shown using the full black lines) clearly show significant deviations from the

[2] Gonzalez-Val (2014) summarized 18 studies published in the last three decennia, ranging from Eaton and Eckstein (1997) to Michaels et al. (2012). Adding to this list three studies that have been published subsequently (Desmet, 2013; Glaeser et al., 2014 and Luckstead and Devados, 2014), there are 11 studies that find evidence supporting Gibrat's law, five that reject it and six showing mixed results.
[3] Studies on Spain, Italy and the United Kingdom not only reject the independent relation for small entities, which is not that uncommon in the literature (Klein, 2014), but also indicate a clear positive relation between growth and size (Klein, 2014; Gonzalez-Val, 2014).

census data. The second panel of this figure shows that the difference between both series can be as high as 15% of total population. The likely causes for this discrepancy are errors in data registration and collection, particularly in the *mouvement* data. Moreover, Vrielinck (2013) points out that municipalities also had a financial incentive to underreport emigration outflow, as more populous cities received more funds from the federal government.

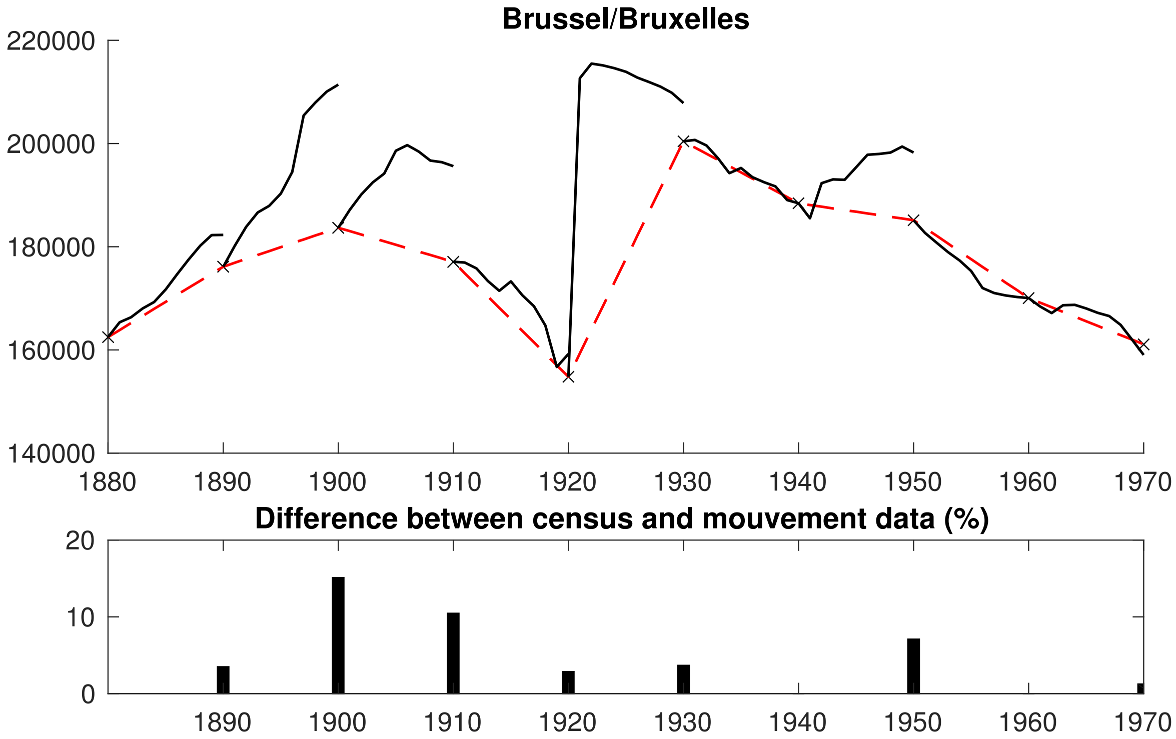

Figure 1 - Population of Brussels based on an amalgamation of ten-yearly census data and yearly population registers (*mouvement*)

Notes: The census data is represented by the black crosses, while the red dotted line shows its imputed values. The black line shows the combination of the census data with the yearly *mouvement* data. The difference between the latter and the census data is shown in the bar graph below, expressed as a percentage of the census data.

While results for Brussels might lead to the conclusion that both series should not be combined, it should be noted the severity of this problem differs depending on the municipality. For example, figure two shows the same plots for the city of Wavre in Wallonia, where the average difference between the census and movement data is less than 1.2 percent.

While adding both series to each other does not produce reliable results, this does not mean that the information from both series cannot be combined. Recent collaborations between econometricians and historians have brought to the fore a technique that could be used to this end, namely state-space models. Simply put, state-space models are used to elicit an unobserved signal (e.g. the level of population in each year) from a collection of data sources of varying quality (e.g. the census and *mouvement* data). The main advantage of these models is their flexibility as they can be adjusted to fit many different situations: in this case, the combination of level and change data that varies in both quality and availability.[4]

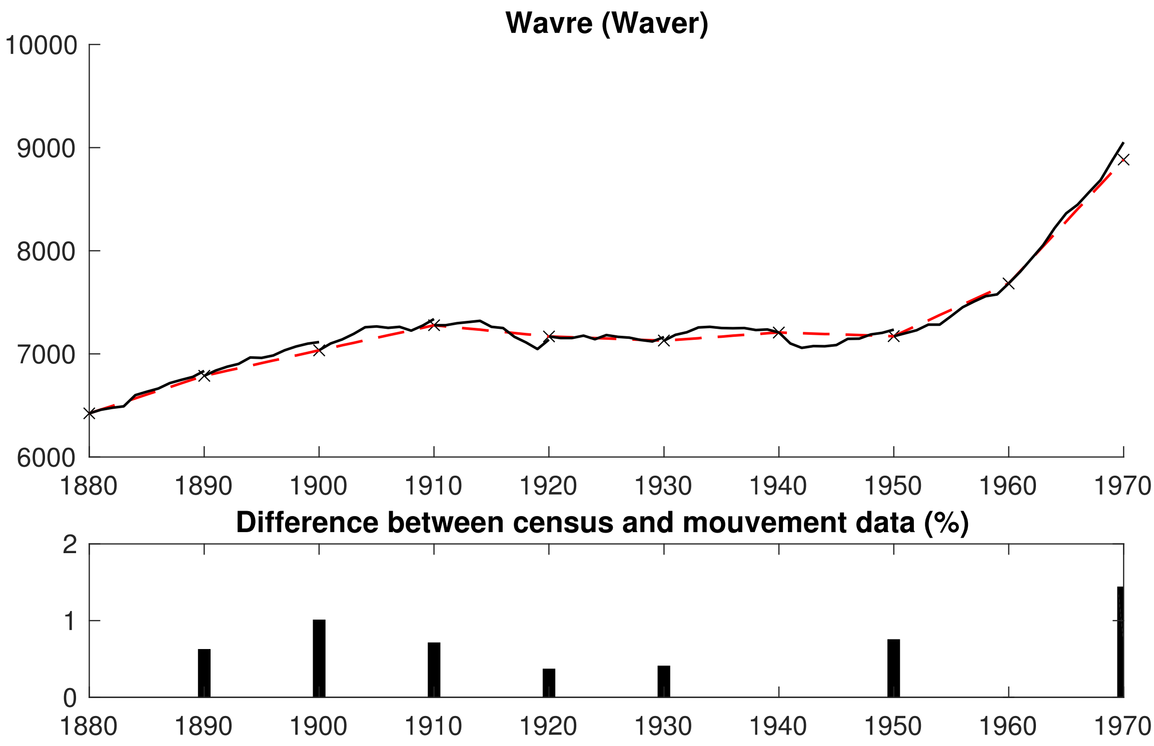

Figure 2 - Population of Wavre based on an amalgamation of ten-yearly census data and yearly population registers (*mouvement*)

Notes: The census data is represented by the black crosses, while the red dotted line shows its imputed values. The black line shows the combination of the census data with the yearly *mouvement* data. The difference between the latter and the census data is shown in the bar graph below, expressed as a percentage of the census data.

The goal of this paper is twofold. We contribute to the growing literature on the application of state-space models in economics. Specifically, we argue that on their flexibility that allows us to combine dataset with markedly different characteristics. Moreover, the assumptions that are used when combining the data can be tested and compared to various alternatives, allowing us to develop a model that maximizes the reliability of the estimates. As a result, this technique allows us to combine different sources of (historical) data, thereby increasing the availability of qualitative historical data. The state-space model also provides estimates of the reliability of its own estimates, as well as that of the underlying sources, allowing these to be taken into account in any subsequent analysis (see e.g. Desbordes and Koop, 2016).

[4] The application of state-space models to solve problems in economic history is relatively recent. The earliest examples explored the use of state-space models in studying Malthusian mechanisms (see e.g. Lee et al., 2002; Crafts and Mills, 2009; or Pfister et al., 2012), while recently Veenstra (2015) and Standaert et al. (2016) used it to combine data on German manufacturing and worldwide bilateral trade flows, respectively.

The second goal of this paper is to produce qualitative estimates of the population of up to 2681 Belgian cities and municipalities on a disaggregated level (LAU level 2) on a yearly basis for the period 1880-1970. The level of detail and the length of the period under analysis should make this dataset of interest to a wide readership, not to mention the fact that this period included both World Wars. This dataset is available upon request.

The remainder of this article briefly discusses the data sources used, after which we give a brief outline of state-space models and explain how they are used to estimate the population levels. After presenting the results, we finish by addressing some critical notes in the conclusion.

2. Data

As mentioned in the introduction, we combine two datasets on the population of Belgian cities and municipalities. The census data was collected at the start of each decade and contains the population levels. The second dataset, called *mouvement,*[5] contains yearly information on the deaths, births, inward and outward migration of each municipality and is based on yearly questionnaires collected by the central government in which each municipality reported on the changes in their population register. Because the data collection of the former was centralized and better controlled, it is of higher quality than the *mouvement* data. The latter are more prone to measurement errors as they were collected in a decentralized manner (both in time and geographically) and errors could have crept in during the initial registration, the collection or transmission of the register data, or could be caused by delays in registration due to administrative procedures. Moreover, the yearly change in population is the result of a combination of four data series (inward and outward migration, births and deaths), further exacerbating the problem. As noted earlier, municipalities also had a financial incentive to inflate their population figures, for example by underreporting outward migration (Vrielinck, 2013).

The datasets were obtained from LOKSTAT[6] that are at present kept at the state archives of Belgium and were disclosed for the period 1880-1976 by the national statistical office (Statistics Belgium). While both contain the same number of municipalities, this number does not stay constant over time. Starting at 2583 in 1880, it continued to rise to 2633 in 1913. The number of municipalities fell back to 2581 in 1919, and while it rose 2670 in 1939, it ended up at 2601 in 1970. Table 2 gives an overview of the data at the beginning and at the end of the studied period. All 2681 distinct municipalities combined, the total number of observations approaches 240.000.

The quality of the dataset is not only dependent on the historical sources, but also on the stability of geographical entities. While the number of municipalities remained relatively stable, a number of

[5] Yearly population figures for the period between 1880 and 1976 coming from the *Mouvement de la Population et de l'Etat Civil* [Population Mouvements and marital status changes], the registers of population, death and the population and the socio-economic censuses

[6] See http://www.lokstat.ugent.be/lokstat_start.php

administrative changes and geographic reorganizations altered the boundaries of the municipalities. These were split up into four types of changes: a full merging of geographic entities, a partial merging of geographic entities, reciprocal border changes between two municipalities and changes in the names of municipalities.

Table 1 - Distribution of population at the beginning and end of the period.

Year	Pop.	Mean	Median	St. dev.	Max	Min
1880	5467783	2163	1060	6790	175636	25
1970	8364488	3820	1291	10187	224545	31

Notes. Overview of total population, mean, median, standard deviation, maximum and minimum population for six representative years. The yearly summary is available upon request.

Finally, the data itself was also checked for large errors. For example, even though no administrative changes were reported, the *mouvement* data suggests that population of Ghent, Antwerp, Liège and Charleroi suddenly increased with half or even ten times the original population. By way of illustration, figure 4 plots the census and mouvement data of Antwerp and Charleroi. However, the census data showed no signs of a change of population of this magnitude in any of these cities during this period, suggesting that either the initial increase was in error or that the corresponding decrease is missing from the dataset. Because of aberrant size of these observations (in some cases, more than 9 times the standard deviation), they tend to skew the estimations if left uncorrected. Looking for similar errors in the dataset, we flagged all observations in the *mouvement* dataset that were bigger than five times the standard deviation for that country. Ignoring those that coincided with administrative changes, as well as those where mouvement and census data matched (within 10% of population), this process identified 43 observations (less than 0.02% of the dataset) as suspect. To improve the workings of the state-space algorithm, these observations were set to missing.

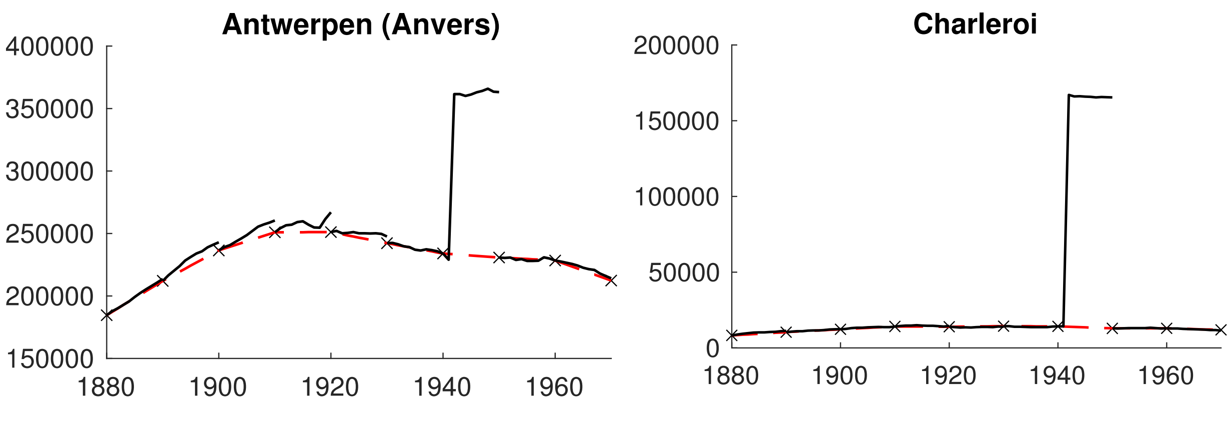

Figure 3 – Major errors in the dataset

Notes: The census data is represented by the black crosses, while the red dotted line shows its imputed values. The black line shows the combination of the census data with the yearly *mouvement* data.

3. State-space population

The following section first discusses the intuition underlying the state-space models and then explains in more detail how the model estimating population was constructed. For more background on state-space models and a thorough explanation on how to estimate them, we refer to Kim and Nelson (1999).

As was noted earlier, a state-space model is a statistical method that enables you to compute the distribution of an unknown variable (the state variable) from a collection of indicators that are related to this state variable in some way. In this case, the unknown state variable is the population of each city, while the measurement variables are the census and *mouvement* data. The way in which these variables are related to the state variable is detailed in the *measurement* equation (2). This equation can be modified to reflect the characteristics of the data, like the fact that the census data captures the level of population, while the *mouvement* data measures the yearly change. Similarly, it can be adjusted to deal with differences in the reliability of the data sources.

A crucial difference with other approaches (e.g. principal component analysis) is that the state-space model also takes the temporal dimension of the data into account. In this case, it is clear that in the absence of large administrative changes or wars, the population of a city strongly depends on that city's population in the previous year. Rather than computing the level of population on a year-by-year basis, the state-space model will use the time-dependence to improve the estimated population level. The way in which population depends on its previous values is detailed in the second equation: the *state* equation (1). As was the case with the measurement equation, the state equation can be adjusted to fit the properties of the data.

If $\boldsymbol{S}_t$ is a vector containing the state variable and $\boldsymbol{X}_t$ is a vector with the measurement variables, a linear state-space typically take the following form:

$$\boldsymbol{S}_t = \boldsymbol{\delta_t} + \boldsymbol{T}\,\boldsymbol{S}_{t-1} + \boldsymbol{\mu}_t \quad with\; \boldsymbol{\mu}_t \sim N(0, Q_t) \tag{1}$$

$$\boldsymbol{X}_t = \boldsymbol{C_t} + \boldsymbol{Z_t}\,\boldsymbol{S}_t + \boldsymbol{\epsilon}_t \quad with\; \boldsymbol{\epsilon}_t \sim N(0, H_t) \tag{2}$$

Vectors $\boldsymbol{C_t}$ and $\boldsymbol{Z_t}$ contain scaling parameters that detail the relation between the state and measurement variables, while the measurement errors are captured by the error term $\boldsymbol{\epsilon}_t$. Vectors $\boldsymbol{\delta_t}$ and $\boldsymbol{T_t}$ details the overall pattern of time-dependence in the state variables, while the error term $\boldsymbol{\mu}_t$ allows the change in the state variable to differ from the overall time-dependence pattern (e.g. because of administrative changes).[7] All of these parameters $\boldsymbol{C_t}, \boldsymbol{Z_t}, \boldsymbol{\delta_t}, \boldsymbol{T_t}, \boldsymbol{Q_t}$ and $\boldsymbol{H_t}$ can vary over time if

[7] Depending on the needs of the model, *C, Z, T, Q* and *H* can all be made time-varying.

necessary, depending on the requirements of the model. In other words, if there is reason to assume that the relationship between the variables changes over time, this can be incorporated in the model.

As illustrated in figure 4, to estimate the value of the state variable in in a certain year, the state-space model will first predict what the current value is using past and future values of the state variable as detailed by the state equation (1).[8] It will subsequently update that prediction using the information from the measurement variables as described in the measurement equation (2). The more reliable the measurement variables are, the stronger the effect of this updating step will be on the value of the state variable, and vice versa.

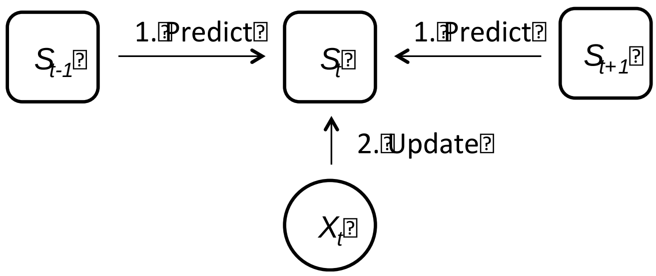

Figure 4 - The estimation of the state-vector

Because the state-space model uses the temporal patterns in the state variable, it is also able to deal with variables that have (many) missing observations. Take the example of the census data that is only available every decade. In the year after the census, the population level will first be predicted using the previous estimate of population (that includes the census data). This estimate will than be updated using the information in the *mouvement* variable. Similarly the year before the census becomes available, the future census data is used to predict what the current population would be. The fact that the census data is unavailable does come at a cost, namely an increase in the uncertainty of the predicted population level. The longer we have to go back (or forward), the larger this uncertainty will be.

3.1. The population state-space model

In order to build the state-space model used to estimate the population of Belgian cities and municipalities, we rely on two main assumptions. In line with the literature (e.g. Vrielinck, 2013), the first assumption that the census data is a correct measure of the actual population, even though it is

[8] To see how future information is taken into account, consider that equation (1) can be rewritten as: $\boldsymbol{S}_t = \boldsymbol{T}^{-1}(\boldsymbol{S}_{t+1} - \boldsymbol{\mu}_t)$

only available every decade.[9] If $P_{i,t}$ is the actual population of city i at time t and $P_{i,t}^{c}$ is its census estimate, this idea can be expressed mathematically as:

$$P_{i,t}^{c} = P_{i,t}. \tag{3}$$

The second assumption is that the yearly change as computed from the city registers (i.e. the *mouvement*) has an error. This *mouvement* data, $\Delta P_{i,t}^{m}$, represent the yearly level change in the population data. This gives us the following mathematical expression for the second assumption:

$$\Delta P_{i,t} = \Delta P_{i,t}^{m} + c_i + \mu_{i,t} \tag{4}$$

The difference between the actual growth rate and the *mouvement* data consists of two parts. $\epsilon_{i,t}$ captures the random measurement errors like those resulting from "honest mistakes" and is normally distributed with mean zero and a city specific variance Σ_i . To allow cities more structural differences, it also contains a city-specific constant c_i that would capture for example consistent underreporting of the outgoing immigration. Because the *mouvement* data is collected and reported by each city individually, the magnitude of both measurement errors is allowed to differ for each city.[10]

In order to build a state-space model from these assumptions, we have to rewrite them such that they fit the shape of the state and measurement equations outlined in equations one and two. To that end, we define equation (3) as the measurement equation and rewrite equation (4) into a state equation. Using the level of population as the state-variable, this gives us the following state-state space model:

$$\begin{aligned} P_{i,t} &= P_{i,t-1} + \Delta P_{i,t}^{m} + c_i + \mu_{i,t} \\ &= \delta_{i,t} + 1 * P_{i,t-1} + \mu_{i,t} \qquad \text{with } \mu_{i,t} \sim N(0, \Sigma_i) \\ P_{i,t}^{C} &= P_{i,t} \\ &= 0 + 1 * P_{i,t} + \epsilon_{i,t} \qquad \text{with } \epsilon_{i,t} \sim N(0,0) \end{aligned} \tag{5}$$

Where $\delta_{i,t} = c_i + \Delta P_{i,t}^{m}$.

3.2. Results

In order to estimate this model, we have to compute the model's parameters ($c_i, \Sigma_i, \Sigma^{\mu}$) and determine the most likely values of the population of all Belgian cities. While this estimation can be achieved using maximum likelihood estimation, we prefer to estimate it using Bayesian Gibbs sampling (with uninformative priors, the results are identical). The latter allows us to split up the estimation problem into more easily solvable subsections, significantly simplifying the problem (see Kim and Nelson,

[9] While assumption is rather restrictive, it is necessitated by the lack of data. The estimations no longer converge when the census data is allowed to have an error, or would require highly restrictive prior assumptions on the variance of the error term.

[10] While imposing the same measurement error for each city ($c_i = c$ and $\Sigma_i = \Sigma$) reduced the number of parameters that have to be estimated with more than 4000, it also decreased the fit of the model to such an extent that the quality of the estimated population went down.

1999). Moreover, the output of the Gibbs sampling estimation is a sample of thousands of draws from the distribution of the population, allowing us to take into account any remaining uncertainty in subsequent analyses or regressions (see e.g. Desbordes and Koop, 2015). For example, this allows us to compute the statistical significance of any changes over time or differences between municipalities.

This algorithm ran for 10,000 iterations and its convergence was verified by inspecting the parameter values.[11] The results for Belgium's four most populous cities are shown in figure 5: Brussels, Antwerp, Ghent and Charleroi. The thick black line plots our estimate of population with its 95% confidence interval indicated by the blue shaded area. For comparison's sake, the red crosses and dotted lines show the census and *mouvement* data, respectively. Finally, the asterisks indicate the years in which administrative changes took place for that municipality.

The graphs in Figure 5 clearly show that the estimated population level tries to follow the pattern of the *mouvement* data, while at the same time ensuring that it equals the census data in each decade. This is the case in those situations where the *mouvement* data does not line up with the census estimates, e.g. the estimates for Brussels in the late 1890s and 1900s or those of Antwerp in the 1920s. Overall, the correlation between both series is very high (95.8%), but this is mostly driven by a very high between correlation (99.94%) i.e. the correlation of the mean values for each country. As we expected, the within correlation, which only compares the changes over time, is much lower (59.7%). What also becomes clear is that the administrative changes are captured by this dataset, like the large increase in the population of Brussels in 1921. Finally, we also see that when the *mouvement* and census data line up better as is the case for Charleroi, the size of the confidence bounds shrinks. In the case of a city like Wavre where match is near perfect, they are practically non-existent.

Finally, regarding the shape of the confidence intervals, you might have expected them to increase as the time to/from the census data increases. That is, as distance with the last certain measurement increases the confidence in the population estimate decreases, resulting in peaks in 1885, 1895, 1905 etc. Instead, the width of the confidence intervals in Figure 5 remains constant in between the census years. The reason is that by setting the population estimate to be exactly equal to the census data in each decade, the total increase over the intervening years is fixed. As a result, more extreme decreases or increases in one year have to be compensated in subsequent years in order to end up at the next census data point. In contrast, for those municipalities where the final census data in 1970 is missing, the confidence bounds increase gradually from 1961 onwards, forming the expected triangular pattern.

[11] Convergence was verified by inspecting their plotted values, autocorrelation functions and CUMSUM graphs.

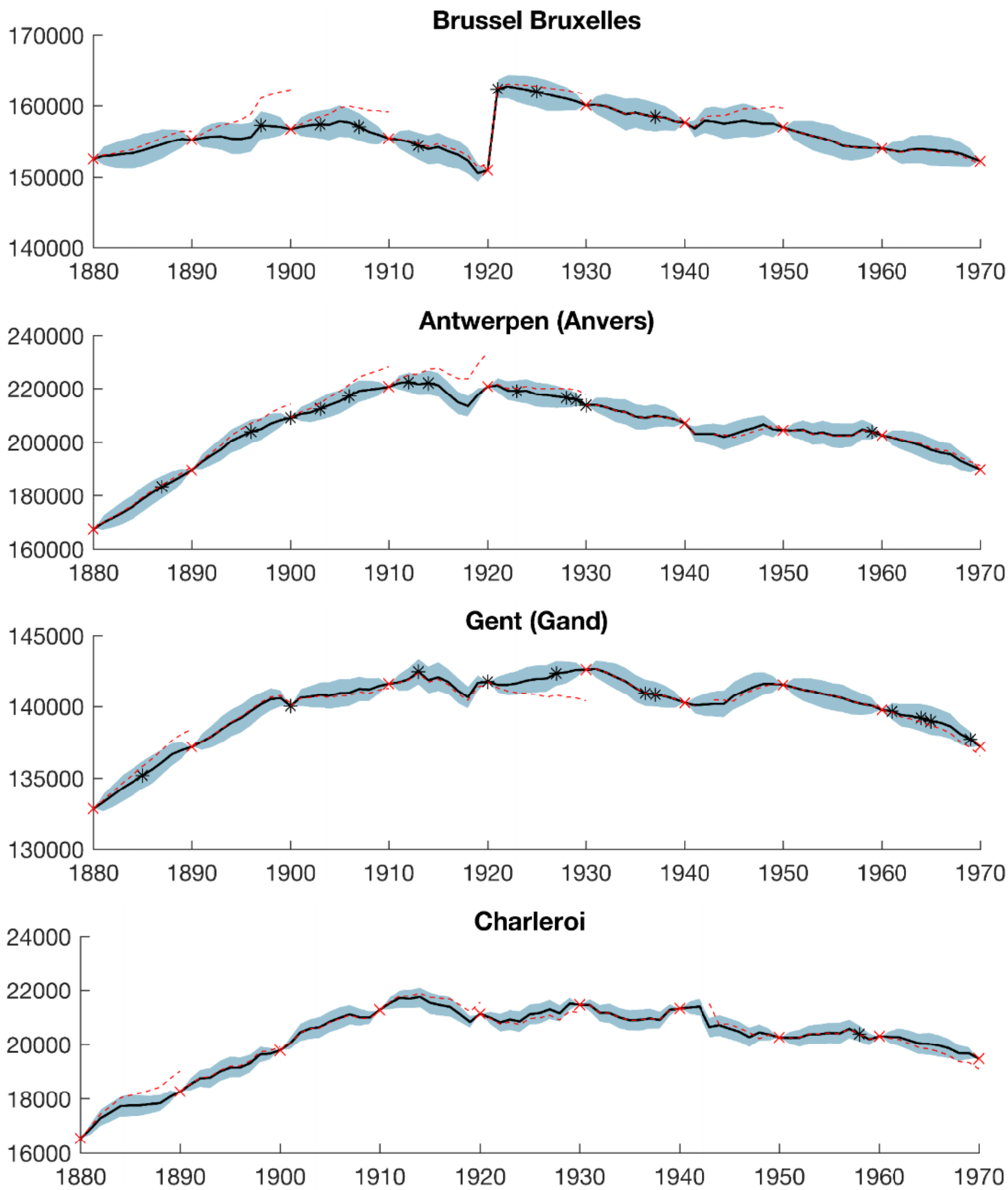

Figure 5 – The state-space population estimates

Notes: The estimated level of population is indicated by the thick black line and its 95% confidence interval by the blue shaded area. The census data and *mouvement* data are plotted using the red crosses and dotted line, respectively. Administrative changes are indicated by asterisk.

Robustness checks

We also tried a number of alternative specifications of the state-space model to see whether these results could be improved, but the specification outlined in equations (3) to (5) outperformed all others. Following Veenstra (2015), we modeled population growth as a local linear trend. This assumes that population growth follows a long-term pattern from which it can deviate in the short term. This model was estimated both in logs and levels, but both versions produced estimates of population that deviated twice as much from the *mouvement* data and had larger confidence intervals.

Secondly, we also considered explicitly incorporating the reported administrative changes into the model. This was done by allowing the variance in the growth rate of population (Σ_i) to be different in these years. While this significantly improved the estimates of the model that used the local linear trend, it did little to improve the estimates using our preferred specification.

Finally, we also considered abandoning the assumption that the census data is without error. However, without the census data as anchor point, the model was unable to produce believable estimates of the level of population. Overall, the relative dearth of data obliges us to impose stronger assumptions in order to produce reliable results.

4. Conclusion

In summary the population data provides a good showcase of the problems that occur when multiple (historical) data series are combined. Not only do both series differ in terms of type of data (levels versus growth) and availability, they also differ in terms of quality. The problem is further compounded by a number of administrative changes that took place throughout the 90-year period. However, this paper has shown that each of these problems can be addressed using state-space models, as it can be fully adjusted to reflect the characteristics of the data that is being combined. As a result, the model can combine both sources of information on population to produce an estimate of the level of population as well as an indicator of its reliability. In this example we construct one of the most replete datasets on population, containing the yearly population growth of over 2600 municipalities for close to a hundred years.

While it is true that a number of assumptions had to be made to construct the state-space model, this also happened when the data is imputed or when the *mouvement* data was added to the census data. For example, all three approaches share the assumption that the census data is correct. However, unlike the alternatives, the state-space model is explicit about which assumptions are made. This transparency allows these assumptions to be more openly critiqued and (hopefully) refined, improving the quality of het data series.

Acknowledgements

We are indebted to Claude Diebolt and Gerdie Everaert for their suggestions and feedback. All remaining errors are of course our own. This project has received funding from the European Union's Horizon 2020 research and innovation program under the Marie Skłodowska-Curie grant agreement No 665501 with the research Foundation Flanders (FWO).

References

Crafts, N. and Mills, T.C., 2009. From Malthus to Solow: How did the Malthusian economy really evolve?. *Journal of macroeconomics*, 31, pp.68-93.

Desbordes, R. and Koop, G., 2015. Should we care about the uncertainty around measures of political-economic development? *Journal of Comparative Economics*, 44, pp. 752-763.

Desmet K. and Rappaport J., 2014. The Settlement of the United States, 1800 to 2000: The Long Transition Towards Gibrat's Law, CEPR Discussion Paper No. DP9353

Eaton, J. and Eckstein, Z., 1997. Cities and growth: theory and evidence from France and Japan, *Regional Science and Urban Economics*, 27, pp. 443–474.

Glaeser E.L., Ponzetto G.A.M. and Tobio K., 2014. Cities, skills and regional change, *Regional Studies*, 48, pp. 7-43

Gonzalez-Val, R., 2014. New evidence on Gibrat's law for cities. *Urban Studies*, 51, pp. 93-115.

Klein A. and Leunig T., 2014. *Gibrat's Law and the British industrial revolution* SERC Discussion Papers, SERCDP0140. Spatial Economics Research Centre (SERC), London School of Economics and Political Science, London, UK

Luckstead J. and Devadoss S., 2014. Do the world's largest cities follow Zipf's and Gibrat's law? *Economic letters*, 125, pp. 182-186

Kim, C.J. and Nelson, C.R., 1999. *State-space models with regime switching: classical and Gibbs-sampling approaches with applications* (Vol. 2). Cambridge, MA: MIT press.

Lee, R. and Anderson, M., 2002. Malthus in state space: Macro economic-demographic relations in English history, 1540 to 1870. *Journal of Population Economics*, 15, pp.195-220.

Michaels, G., Rauch, F. and Redding, S. J., 2012. Urbanization and structural transformation, *The Quarterly Journal of Economics*, 127, pp. 535–586.

Pfister, U., Riedel, J. and Uebele, M., 2012. Real wages and the origins of modern economic growth in Germany, 16th to 19th centuries. EHES working paper (No. 0017).

Standaert, S., Ronsse, S. and Vandermarliere, B., 2016. Historical trade integration: globalization and the distance puzzle in the long twentieth century. *Cliometrica*, 10, pp. 225-250.

Veenstra, J., 2015. Output growth in German manufacturing, 1907–1936. A reinterpretation of time-series evidence. *Explorations in Economic History*, 57, pp. 38-49.

Vrielinck S., 2013. De macht van het getal: een statistische kritiek van de bevolkingsstatistieken in België (1801-1976) [The power of numbers: a statistical critique on population censuses in Belgium (1801-1976)]. *Journal of Belgian History - Belgisch Tijdschrift voor Nieuwste Geschiedenis*, 43, p. 74-113.