Gaussian Process Regression for Analysis of the Inbound Commuting to the Netherlands

Abdulhakim Özcan
Koç University

Abstract

The European Union (EU) has taken enormous steps towards integration since its establishment. The creation of a common market in the 1990s hugely contributed to this process. One of the benefits enjoyed in the EU common market is to work in one member state while living in another. The EU also implements a diverse set of policy instruments to ensure inclusion for all. In this paper, I want to examine the inclusiveness of the EU's economic integration process across the Dutch borders, analyzing the gender distribution of inbound commuters from Belgium and Germany. I will endeavor to model the over-time distribution of female employees of 3 different nationalities who commutes to 40 Dutch regions at NUTS 3 level, utilizing Gaussian Process Regression (GPR).

Introduction

The EU is one of the world pioneers in many different fields from economic and democratic development to sustainability. The region is also very visible with its stance on gender equality which is indeed one of the union's founding values. According to World Economic Forum (WEF, 2021), Western Europe continues to progress the most towards gender equality with the highest overall score of 77.6%. However, the region lags behind two other regions –North America (75.3%) and Eastern Europe and Central Asia (73.8%)– with its gender gap of 70.0% on the Economic Participation and Opportunity subindex (WEF, 2021). Moreover, it seems that the gap widens further among cross-border commuters in certain countries. For example, 32% of the workers commuting to Germany were Polish men employed in the construction field, whereas 46% of workers commuting to Austria were Slovakian women employed in human health (Eurostat, 2019). However, the Netherlands has a narrower gender gap compared to the majority of Western Europe with a women/men ratio of 0.90 in labor force participation (WEF, 2021). However, the same ratio might be lower among the Belgian and German workers commuting to the Netherlands, considering that there may be additional barriers when working outside the country of residence. I may claim that the over-time percentage of women among these commuters is significantly lower. Additionally, I may expect that the Belgian and German female commuters show a resembling over-time trend in terms of their percentage among other incoming commuters.

Data

The EU defines cross-border commuters as workers who are employed in one member state while officially living in another and return to the country of residence daily or at least once a week (EU, n.d.). In 2010, almost 92 thousand people were employed in the Netherlands while they were living in Belgium or Germany whereas this number dropped to over 82 thousand people in 2018 (Grensdata, 2021). These cross-border commuters are approximately evenly distributed between the Belgian and German municipalities just across to the Dutch border (see **Exhibit 1**). The commuting data used to feed into the model is received from the Netherlands' national statistical office (CBS). The office identified the workplace of the commuters for the first time in 2020 and the data is made available with a period ranging from 2010 to 2018. CBS provides border statistics in a separate database which is available only in German and Dutch. To understand and read the website, I used Chrome's built-in translation feature to translate the web pages to English. Similarly, I used translation to understand the downloaded data and metadata. CBS provides the data on the aggregate level as well as on a more cross-sectional level. For example, the annual number of commuters can be broken down by sex, nationality, working regions (NUTS levels), and country of residence and the office allows people to download the data based on various combinations of these categories. The year and two other categories -sex and country of residence- are must for this paper as it intends to analyze the over-time gender distribution of incoming commuters. However, I decided to break down the data further into nationality and working regions to be able to test all the hypotheses while, at the same time, hope to catch NUTS 3 level variances. The final data contains two different countries of residence –Belgium and Germany– three different nationalities –Dutch, Belgian, and German– and 40 different NUTS 3 regions. The dataset allows tracing the workers based on their sex, nationality, country of residence, and country of the workplace over the given period. For example, I can see the distribution of Belgian female workers commuting to the Netherlands from 2010 to 2019. Whether I break the data down into further categories or download it with two main categories, it fits into the definition of panel data where different units are observed over a period (Research HUB, 2019).

Method

The longitudinal panel data fed into this model is obtained from 40 different statistical COROP regions of the country at 9 time points from 2010 to 2018. It seems that each COROP region has had some commuters from Belgium or Germany at some point in time. Due to the complex nature of the data, I used the Gaussian process (GP) which is a flexible Bayesian non-parametric method to analyze underlying function (Rasmussen & Williams, 2006). The motivation behind the

GPs is that the default approaches are often not suited to model noisy, large-scale, and non-linear real-world data sets.

$$f(x) \sim GP\big(m(x), k(x, x')\big) \tag{1}$$

GP has a mean function and a covariance function as seen in (1), rather than having a mean and a covariance matrix (Rasmussen & Williams, 2006, p. 13). The first expression on the right side of the equation is called the mean function $m(x)$, while the second expression is the covariance function. Covariance functions receive certain inputs, x and x', and essentially produce covariance matrices corresponding to those entries. To construct the covariance function, I used the Exponentiated Quadratic kernel which is also referred to as the Squared Exponential or Radial Basis Function kernel. The Exponentiated Quadratic kernel is known for modeling the smoothness of function and it is also utilized to encode the local variation within a dataset. Considering that my data includes closely located regions and a time component, this kernel seem to be a useful kernel for my purpose. Indeed, this kernel performed either better or seemingly similar compared to a few other kernels. On the other hand, mean functions, which are regarded as less important than specifying the covariance function, create mean vectors and they are usually set to constant values or linear functions (Fonnesbeck, 2020). Like other Bayesian models, GP is specified based on the Bayesian formula. Equation (2) shows the way the Bayesian model is specified (Steorts, n.d.).

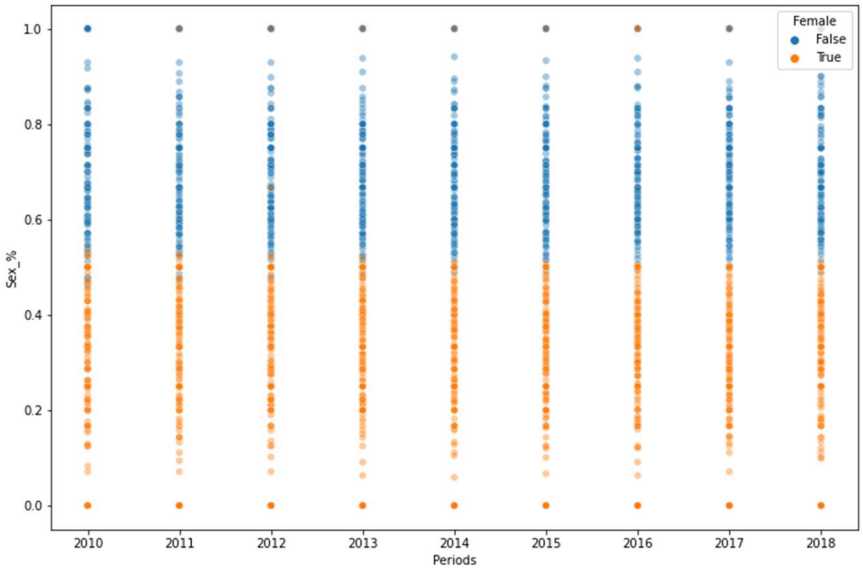
$$P\big(\theta \mid x\big) \propto P\big(\mathrm{x} \mid \theta\big)\, P(x) \tag{2}$$

The expression on the left side of the formula is called posterior and it represents the probability of "θ", unknown, given some observed features "x", data (VanderPlas, 2016). The expressions on the right side of the formula are respectively called likelihood function and prior distribution and both together represent the combination of information about the "θ" before the data is observed with information from the data (Fonnesbeck, 2020). In the Gaussian process, we are essentially interested in estimating the posterior distribution, the left side of (2). To implement a Gaussian process regression for the commuting data mentioned earlier, the same Bayesian steps were followed. There are multiple libraries to implement GP in Python. GPy, GPflow, PyStan TensorFlow Probability, and scikit-learn are some of these open-source libraries. PyStan and TensorFlow Probability are considered as low-level toolkits which are closer to machine language whereas others are considered higher-level libraries to varying degrees (Fonnesbeck, 2020). This paper will use PYMC3 to define and use GPs. PYMC3 is an open-source high-level probabilistic programming package written into Python that allows the construction and implementation of Bayesian models (Salvatier et al., n.d.). While running the code, you may get divergences but their numbers are around 5-6 and it seems that this number is not a terrible problem (Fonnesbeck, 2020)

Findings

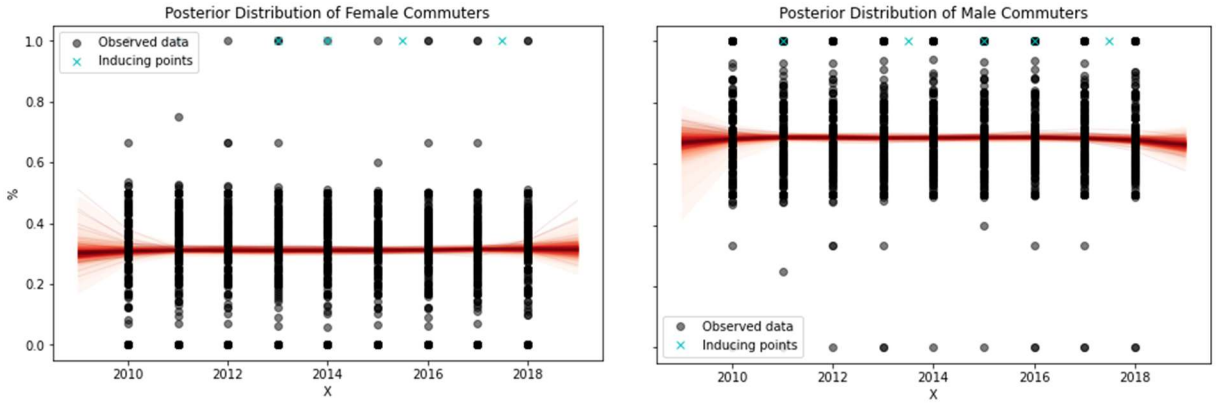
I mentioned the women/men ratio earlier in the introduction. This study did not investigate the women/men ratio but rather the percentage of women among all workers. The women/men ratio is intentionally not investigated because some COROP regions in the Netherlands show complete domination by either of the genders. To avoid zero divisions and not waste potential information, I looked at the percentage of women in each region. **Figure 1** shows the sex distribution of inbound commuters from Belgium and Germany.

***Figure 1**: Percentage of Commuters by Gender (2010-2018)*



I was not very optimistic about the gender distribution among the commuters, and it seems that people commuting to the Netherlands are dominated by male workers. However, there still are some regions that females with certain nationality and country of residence dominate the region. According to the data, almost 68.6 % percent of the workers commuting to the Netherlands over the 9 years were male while 31.4 % of the workers were female.

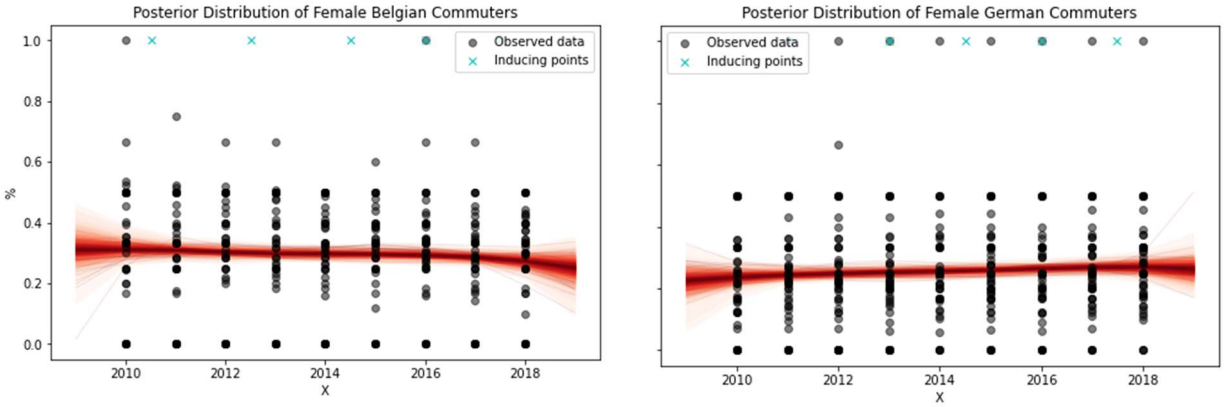
***Figure 2**: Posterior Distribution of Commuters by Sex (2010-2018)*



According to the fitted Gaussian process model, female commuters have shown a very constant linear trend around 30 % from 2010 to 2018. On the other hand, male commuters show a constant trend of around 70 % over the same period. It should be noted that the is some uncertainty on the edges because there is no data before 2010 and after 2018. The first hypothesis suggested that the over-time percentage of women among incoming commuters would be significantly dominated by male commuters. However, almost one-third of the commuters are female for 9 years and I may claim that this number is not significantly low. Thus, I may consider the first hypothesis as defeated.

Based on no theoretical framework or information, the second hypothesis expected Belgian and German females to show a resembling over-time trend in their percentage among the inbound commuters. The GP model shows that these groups show different trends (**Figure 2**). While the overall percentage of female Belgian commuters in their respective COROP regions decreased over time, the same ratio increased for female German commuters.

***Figure 3**: Posterior Distribution of Commuters by Nationality (2010-2018)*



Limitations

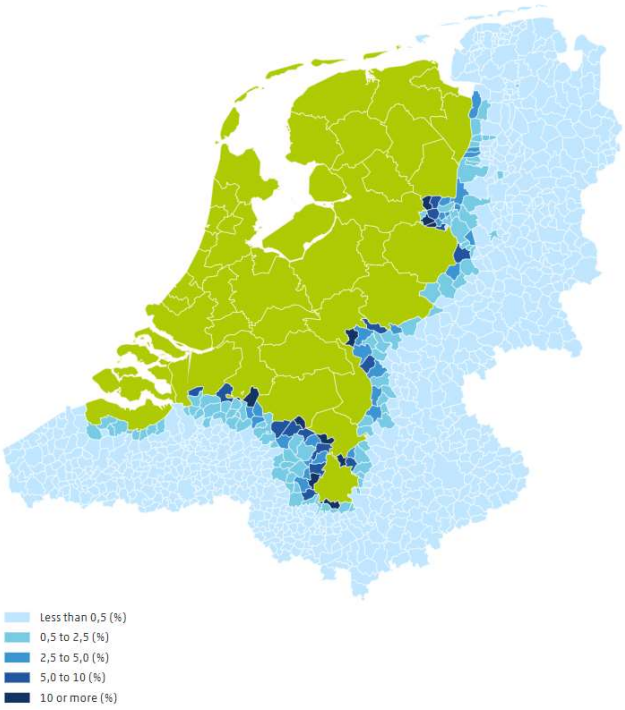
It seems that the Gaussian process model is not one of the easiest and intuitive models to learn. The model is abstract and harder to understand and implement compared to other machine learning algorithms that I have learned so far. On the other hand, there seem to fewer online resources dedicated to the PYMC3 implementation of this model. I spent a good amount of time searching for a way to implement a multi-output Gaussian process that simultaneously models the over-time distribution of different populations in the same process, but I failed. There are several online implementations of the multi-output Gaussian process in PYMC3. I tried to adopt the ideas behind them into the source code used to create this model, but these implementations often lacked the time component, and it was hard to find commonalities between them and this model. That is why I had to plot the distribution of female and male, or female Belgian and German commuters separately. The overall process was somewhat challenging considering both the nature of the data and the model.

Conclusion

The specified Gaussian process model defeated both of my hypotheses. However, the model may prove to be useful for decision-makers in the EU and the Netherlands. Inclusiveness is an important component of the EU policies it may need further attention in certain countries or regions at different levels. Although the percentage of female workers commuting to the Netherlands is not substantially low, NUTS 3 regions in the country are dominated by male commuters. This model could potentially help the decision-makers to analyze the over-time distribution of commuters by sex, nationality, country of residence, and region of the workplace. However, the model has a lot of room for improvement.

Appendix:

***Exhibit 1**: Municipal Share of Inbound Cross-Border Commuters (2018)*



Source: CBS. (2020, November 18). *Inbound commuters often live just across the border*. https://www.cbs.nl/en-gb/news/2020/46/inbound-commuters-often-live-just-across-the-border

Reference:

Coding Tech. (2020, January 8). *A Primer on Gaussian Processes for Regression Analysis*. YouTube. https://www.youtube.com/watch?v=xBE8qdAAj3w&t=2947s&ab_channel=CodingTech

EU. (n.d.). *Cross-border commuters - Your Europe*. Retrieved June 7, 2021, from https://europa.eu/youreurope/citizens/work/work-abroad/cross-border-commuters/index_en.htm

Eurostat. (2019). *2.3 Crossing borders*. https://ec.europa.eu/eurostat/cache/digpub/eumove/bloc-2c.html?lang=en

Grensdata. (2021, February 12). *Grenspendel werknemers; nationaliteit, woonland, werkregio (NUTS 3)*. https://grensdata.eu/#/InterReg/nl/dataset/22003NED/table?ts=1622891052614

Rasmussen, C. E., & Williams, C. K. I. (2006). Gaussian processes for machine learning LK - https://kocuniversity.on.worldcat.org/oclc/61285753. In *Adaptive computation and machine learning TA - TT* -. MIT Press. http://catdir.loc.gov/catdir/toc/fy0614/2005053433.html

Research HUB. (2019, August 6). *Panel Data (1): Introduction to Panel Data Analysis*. YouTube. https://www.youtube.com/watch?v=drxZhzdyj-g&ab_channel=ResearchHUB

Salvatier, J., Wiecki, T. V., & Fonnesbeck, C. (n.d.). *Getting started with PyMC3 — PyMC3 3.10.0 documentation*. Retrieved June 9, 2021, from https://docs.pymc.io/notebooks/getting_started.html

Steorts, R. C. (n.d.). *Module 1: Introduction to Bayesian Statistics, Part I*. Retrieved June 9, 2021, from http://www2.stat.duke.edu/~rcs46/modern_bayes17/lecturesModernBayes17/lecture-1/01-intro-to-Bayes.pdf

VanderPlas, J. (2016). *Python Data Science Handbook: Essential Tools for Working with Data* (1st ed.). O'Reilly Media, Inc.

World Economic Forum. (2021). *Global Gender Gap Report 2021*. http://www3.weforum.org/docs/WEF_GGGR_2021.pdf