Searching for the best traffic data to predict Covid-19 infections in Nordic countries

Tommi Gröhn

Abstract

Introduction: When an epidemic starts to escalate in a specific region, people should reduce their traffic to get the spread of the disease under control again. With a data-driven approach this study retrospectively answers to the question: What kind of people's traffic data country officials should have followed when the second wave hit Scandinavian counties?

Methology: For 9 different regions in Finland, Norway and Sweden, we collect data on Covid-19 infections and people's traffic behavior. We build a stochastic SEIR-type model where we add different Google's traffic components into the model. For each region, we use leave future out -methology to evaluate the goodness of a fit to find out which traffic component gave the best fit.

Results: Models where transit station or residential traffic were used performed the best in predicting the future infections in a region.

Discussion: In most regions we got reasonable parameter estimates and predictions but our model did not describe well the escalation of the epidemic in Pirkanmaa or Stockholm for any traffic component. Our results suggest that transit station traffic data is the best Google's traffic component to follow by country officials when making predictions of future infections.

I. Introduction

We retrospectively study what kind of Google's traffic data was the best to predict Covid-19 infections in Scandinavia during the second wave [1]. Should country officials have followed people's retail, grocery, park, transit station, workplace or residential traffic when making predictions of new infections? Of course our approach studies only one wave of one disease in 9 regions but this approach still gives some advice for country officials on which kind of traffic data they should build their predictions on in the future epidemics.

We assume people's traffic behavior strongly correlates with the spread of Covid-19. Our analysis is based on open data available for three counties in Finland, Norway and Sweden

each: Pirkanmaa, Southwest Finland (Varsinais-Suomi), Uusimaa, Oslo county, Vestland, Viken, Skåne, Stockholm county and Västra Götaland.

We created a tool for our analysis which is based on a stochastic SEIR-type model implemented by Norwegian Institute of Public Health [2]. In general, SEIR-type models have been commonly used to describe Covid-19 pandemic [3], [4]. However our model differs from the others by adding in the model data on people's traffic behavior. For each region, we fit the described model to the reported statistics of Covid-19 infections and use the resulting parameter values as features characterizing a specific Google's traffic component. We evaluate the goodness of a fit using leave-future-out methology using STAN-libraries and -documentations [5], [6]. Our model makes explicit the role of diagnostic and reporting delays.

II. Methology

A. Data

We collected the infected data from THL Sampo's, Norwegian Institute of Public Health's and Folkhalsomyndigheten's open data sources [7]–[9]. We remove a clear outlier on Southern Finland's infection data on date 30.11.2020 and use the average value of infections on 29.11.2020 and 1.2.2020 instead.

Furthermore, 6 different traffic components from each region was collected from Google's open data. These six components are retail & recreation, grocery & pharmacy, parks, transit stations, workplaces and residential. Google has counted a baseline value for each traffic component and for each region which is counted as a median value of multiple days in February. If a traffic component i's value is -20, it means that a region has reduced 20% of its traffic on that current date compared to the value in February.

B. Choosing suitable time periods

During the spring 2020, Covid-19 caused massive lockdowns in European countries, including Finland and Norway. Therefore it is difficult to analyse how the pandemic was got under control during the spring. Was it the effect of people spending less time in supermarkets or at transit stations?

However, people slowly started to change their traffic behaviour towards normal during summer 2020. We assume this slow change in people's traffic behavior eventually reached the point that the epidemic started to escalate again during autumn 2020. Furthermore, when the epidemic

was got under control again it was done during the autumn with smaller traffic changes than in spring 2020. Therefore we concentrate our analysis on autumn 2020 before the Christmas, i.e. the second wave.

We take a weekly sum of infections to reduce the uncertainty of infections of one single day. We identify the biggest ratio between weekly infections during autumn 2020, i.e. September, October and November. For example if there were 500 weekly infections after a week with 200 infections, the corresponding ratio would be 2.5. We denote the investigated infections with $\{\hat{I}_0, \hat{I}_1, \ldots, \hat{I}_9\}$. The fraction $\frac{\hat{I}_1}{\hat{I}_0}$ corresponds the biggest ratio during autumn which is a logical choice for starting point. Indeed, the ratio usually gets later on smaller because people reduce their traffic behavior after the starting point. $\hat{I}_0$ is used only for initializing the model. The actual training and test period are then subsets of $\{\hat{I}_1, \ldots, \hat{I}_9\}$.

We calculate a weekly average of each traffic component in a way that there is a three day lag between the traffic weekly average and the weekly sum of infections. Then we create to each vector of weekly infections a corresponding vector with 10 elements of the weekly average for each traffic component i. For clarity, we sum each element with a constant s.t. the first element of the vector is 0 and then we remove the first element. Indeed, our created vector for traffic component i can be described with $\{\tau_1^i, \tau_1^i, \ldots, \tau_9^i\}$ where the element $\tau_0^i = 0$ is not included in the vector.

For each region we have plotted $\{\hat{I}_0, \hat{I}_1, \ldots, \hat{I}_9\}$ and $\{\tau_1^i, \tau_1^i, \ldots, \tau_9^i\}$ in the section VI.

C. SEAPIR-model

General description:

We fit the following Bayesian model for each region separately.

We represent individual regions as data points which are characterized by Covid-19 pandemic related statistics. Our model follows a stochastic SEIR-type model implemented by Norwegian Institute of Public Health where a population N is divided in six groups: susceptibles S, exposed E, infected I, asymptomatic cases A, hospitalized H and removed R. Notice in the asymptomatic group A may belong people with minor symptoms if these symptoms are not reported further to public officials. On contrary, in the group infected I belong only those cases which are officially reported. The group removed R includes everyone who is either recovered from the disease or died because of it. The changes of these groups can be described by the following equations:

$$\frac{\partial S}{\partial t} = -(\frac{r_A \beta S A}{N} + \frac{r_P \beta S P}{N} + \frac{\beta S I}{N}) \tag{1}$$

$$\frac{\partial E}{\partial t} = \frac{r_A \beta S A}{N} + \frac{r_P \beta S P}{N} + \frac{\beta S I}{N} - \frac{E}{D_e} \tag{2}$$

$$\frac{\partial A}{\partial t} = \frac{rE}{D_e} - \frac{A}{D_a} \tag{3}$$

$$\frac{\partial P}{\partial t} = \frac{(1-r)E}{D_e} - \frac{P}{D_p} \tag{4}$$

$$\frac{\partial I}{\partial t} = \frac{P}{D_p} - \frac{I}{D_i} \tag{5}$$

$$\frac{\partial R}{\partial t} = \frac{I}{D_i} + \frac{A}{D_a} \tag{6}$$

We use a week as a time unit and use the same parameter values as FHI: $D_e = \frac{3}{7}$, $D_p = \frac{2}{7}$, $D_i = \frac{5}{7}$, $r = 0.4$, $r_a = 0.1$ and $r_p = 1.25$.

However, we don't estimate β directly because then the created SEIR-model would only be able to describe exponential growth. In reality the growth of infections can also be for example linear as people change their behavior to get the epidemic under control.

Therefore we formulate

$$\beta(t) = c_1 + \tau_i(t) c_2 \tag{7}$$

where $\tau_i(t) = \tau^i_{\lceil t \rceil}$ is a step function for traffic component i.

Prior and likelihood choices:

Indeed, we estimate c_1 and c_2 with our Bayesian model. We use weakly informative priors $c_1, c_2 \sim \text{Normal}(0, 10)$.

Furthermore, we fit our model to weekly summed infected data.

$$\hat{I}_t \sim \prod_t \text{NegBinomial}(\lambda_t, \phi) = \prod_t \binom{y + \phi - 1}{y} \binom{\lambda_t}{\lambda_t + \phi}^y \binom{\phi}{\lambda_t + \phi}^\phi \tag{8}$$

where $\lambda_t = \frac{P(t)}{D_p}$, $t \in \{1, \ldots, 9\}$ and $\frac{1}{\phi} \sim \exp(5)$.

Initialization:

We initialize the model recursively using information of $\hat{I}_0$ s.t.

- $P(0) = \hat{I}_0 D_p$
- $E(0) = \frac{1}{r}\frac{D_e}{D_p}P(0)$
- $I(0) = (1-r)\frac{D_i}{D_e}E(0)$
- $A(0) = r\frac{D_i}{D_e}E(0)$
- $R(0) = 0$
- $S(0) = N - I(0) - P(0) - A(0) - E(0) - R(0)$

D. Validation

We will validate our model using leave-future-out methology. We choose $L = 5, M = 2$. Indeed, we get $p(y_{i+1}, y_{i+2} \mid y_1, \ldots, y_i) = \frac{1}{S}\sum_{s=1}^{S} p(y_{i+1}, y_{i+2} \mid y_1, \ldots, y_i, \theta_{1:i}^{(s)})$ for each $i \in \{5, 6, 7\}$ where $S = 4000$ is the sample size. Eventually we evaluate each traffic component using formula

$$\sum_{i \in \{5,6,7\}} \ln\, p(y_{i+1}, y_{i+2} \mid y_1, \ldots, y_i) \tag{9}$$

Furthermore, we collect 95% -intervals for c_1 and c_2.

We also have visualisations of posterior predictions in our Jupyter notebook.

III. Results

TABLE I: LFO-values (best values highlighted with grey-blue)

	Retail	Grocery	Parks	Transit stations	Workplace	Residential
Pirkanmaa	−36.6	−36.4	−37.0	−37.3	−34.5	−35.8
Southwest Finland	−32.9	−33.5	−34.0	−33.2	−35.3	−33.2
Uusimaa	−40.2	−44.2	−45.0	−39.4	−45.7	−40.4
Oslo	−45.3	−49.9	−49.0	−41.2	−41.0	−41.9
Vestland	−46.0	−45.5	−40.2	−38.3	−40.3	−39.4
Viken	−57.5	−53.8	−50.6	−40.9	−48.1	−46.5
Skåne	−52.4	−55.0	−53.2	−52.4	−54.7	−52.1
Stockholm	−57.8	−56.1	−56.6	−57.1	−55.6	−56.9
Västra Götaland	−52.0	−54.1	−54.1	−54.0	−54.4	−52.7
Σ	−421	−429	−420	−394	−410	−399

We observe from table I our model gives the best predictions when transit station or residential traffic is used. Neither of these traffic components had a county where they performed significantly worse than any other traffic component. Furthermore, transit station traffic made the best predictions in Uusimaa, Vestland and Viken. Residential traffic made the best predictions in Skåne.

Retail and workplace traffic made good predictions in some regions but performed quite badly in many regions. Grocery and park traffic gave in general bad predictions.

We observe from the table III that in Pirkanmaa, Southwest Finland and Stockholm, c_2 got both positive and negative in its posterior. Indeed, in these regions it was unclear whether an increace in a specific traffic component made β down or up.

All the R-hat-values for different variables are closer 1 than 1.05 which is a commonly used diagnostic that the chains have converged.

IV. Discussion

Based on LFO-values, we conclude that country officials should follow transit station or residential traffic when predicting future infections. Practically it may be easier to follow public transport hubs such as subway, bus, and train stations than people's residential traffic.

It is intuitive that grocery or park traffic does not seem to be important kind of traffic to follow. It is often easy to keep distance to other people outside and for example in supermarkets.

It may surprising retail and workplace traffic made quite bad predictions in general. Indeed, a reduction in these traffic component did not indicate a reduction of infections in the future. It is good to emphasize that reducing retail and workplace traffic may have played a significant role in the fight against the Covid-19-pandemic. These traffic components just do not seem to be the best ones to follow by country officials when a new wave of pandemic hits a specific region.

Pirkanmaa and Stockholm were regions where transit station traffic performed badly. However these regions were the most problematic ones in general because none of the traffic components learned c_2 to be either positive or negative. Indeed, in these regions all the traffic components performed badly. Our choice for the investigated time period maybe should have been different for these regions.

Indeed, our results suggest that out of all Google's traffic components transit station traffic described best the spread of the second Covid-19 wave in Scandinavia.

References

[1] Google. (2021) Covid-19 community mobility reports. [Online]. Available: https://www.google.com/covid19/mobility/

[2] B. F. D. Blasio, R. A. White, A. D.-L. Palomares, G. M. Grøneng, J. C. Lindstrøm, M. N. Osnes, F. D. Ruscio, A. B. Kristoffersen, and G. Øyvind Isaksson Rø, "Situational awareness and forecasting for norway," 2021. [Online]. Available: https://www.fhi.no/contentassets/e6b5660fc35740c8bb2a32bfe0cc45d1/vedlegg/nasjonale-og-regionale-rapporter/national_regional_model_17_february_2021.pdf

[3] C. Wang, L. Liu, X. Hao, H. Guo, Q. Wang, J. Huang, N. He, H. Yu, X. Lin, A. Pan, S. Wei, and T. Wu, "Evolving epidemiology and impact of non-pharmaceutical interventions on the outbreak of coronavirus disease 2019 in wuhan, china," *JAMA*, 2020. [Online]. Available: https://www.medrxiv.org/content/10.1101/2020.03.03.20030593v1.full.pdf

[4] A. Hauser, M. J. Counotte, C. C. Margossian, G. Konstantinoudis, N. Low, C. L. Althaus, and J. Riou, "Estimation of sars-cov-2 mortality during the early stages of an epidemic: A modeling study in hubei, china, and six regions in europe," *PLOS Medicine*, vol. 17, no. 7, pp. 1–17, 07 2020. [Online]. Available: https://doi.org/10.1371/journal.pmed.1003189

[5] L. Grinsztajn, E. Semenova, C. C. Margossian, and J. Riou, "Bayesian workflow for disease transmission modeling in stan," 2021. [Online]. Available: https://arxiv.org/pdf/2006.02985.pdf

[6] P.-C. Bürkner, J. Gabry, and A. Vehtari, "Approximate leave-future-out cross-validation for bayesian time series models," *Journal of Statistical Computation and Simulation*, vol. 90, no. 14, p. 2499–2523, Jun 2020. [Online]. Available: http://dx.doi.org/10.1080/00949655.2020.1783262

[7] THL. (2021) Covid-19 cases in the infectious diseases registry. [Online]. Available: https://sampo.thl.fi/pivot/prod/en/epirapo/covid19case/fact_epirapo_covid19case?&row=hcdmunicipality2020-445222&column=dateweek20200101-509030

[8] FHI. (2021) Daily report and statistics about coronavirus and covid-19. [Online]. Available: https://www.fhi.no/en/id/infectious-diseases/coronavirus/daily-reports/daily-reports-COVID19/

[9] ArcGIS. (2021) Folkhalsomyndigheten covid19. [Online]. Available: https://www.arcgis.com/sharing/rest/content/items/b5e7488e117749c19881cce45db13f7e/data

V. Appendix: Confidence intervals for c_1 and c_2

TABLE II: 95% -confidence interval for c_1

	Retail	Grocery	Parks	Transit stations	Workplace	Residential
Pirkanmaa	$[1.4, 1.9]$	$[1.5, 1.9]$	$[1.6, 2.4]$	$[1.5, 2.8]$	$[1.5, 2.1]$	$[1.6, 2.4]$
Southwest Finland	$[1.7, 2.4]$	$[1.7, 2.2]$	$[1.6, 2.4]$	$[1.5, 2.8]$	$[1.5, 2.1]$	$[1.6, 2.4]$
Uusimaa	$[2.0, 2.3]$	$[1.8, 2.1]$	$[2.0, 3.3]$	$[1.9, 2.1]$	$[1.5, 2.1]$	$[2.0, 2.3]$
Oslo	$[1.9, 3.2]$	$[1.4, 1.8]$	$[0.8, 3.7]$	$[2.2, 3.0]$	$[2.1, 2.7]$	$[2.1, 2.9]$
Vestland	$[1.0, 2.5]$	$[1.3, 1.8]$	$[2.6, 7.0]$	$[2.5, 4.2]$	$[1.8, 3.0]$	$[1.8, 3.2]$
Viken	$[1.4, 2.5]$	$[1.8, 2.1]$	$[2.0, 4.7]$	$[3.0, 3.7]$	$[2.4, 3.3]$	$[2.6, 3.8]$
Skåne	$[1.0, 2.5]$	$[1.3, 1.8]$	$[2.6, 7.0]$	$[2.5, 4.2]$	$[1.8, 3.0]$	$[1.8, 3.2]$
Stockholm	$[2.1, 2.2]$	$[1.9, 2.2]$	$[1.9, 2.2]$	$[2.0, 2.2]$	$[2.0, 2.6]$	$[2.0, 2.2]$
Västra Götaland	$[2.3, 2.6]$	$[2.2, 2.4]$	$[2.2, 2.5]$	$[2.3, 2.6]$	$[2.1, 2.8]$	$[2.3, 2.6]$

TABLE III: 95% -confidence interval for c_2 (yellow: both negative and positive values)

	Retail	Grocery	Parks	Transit stations	Workplace	Residential
Pirkanmaa	$[-15, 8.4]$	$[-17, 8.8]$	$[-1.6, 0.5]$	$[-12, 6.0]$	$[-3.7, 9.4]$	$[-9.4, 18]$
Southwest Finland	$[-2.2, 14]$	$[-4.3, 15]$	$[-0.3, 1.0]$	$[-1.7, 6.0]$	$[-9.5, 6.4]$	$[-16, 4.6]$
Uusimaa	$[7.8, 17]$	$[-4.3, 31]$	$[0.2, 1.5]$	$[3.6, 11]$	$[-4.0, 11]$	$[-20, -3.8]$
Oslo	$[2.2, 14]$	$[-14, 4.6]$	$[-3.1, 6.4]$	$[5.5, 14]$	$[4.9, 11]$	$[-22, -8.0]$
Vestland	$[-6.0, 8.2]$	$[-5.4, 3.5]$	$[2.5, 12]$	$[5.1, 13]$	$[5.4, 22]$	$[-34, -6.1]$
Viken	$[-8.3, 7.0]$	$[-7.1, 1.1]$	$[0.1, 5.1]$	$[9.9, 16]$	$[9.0, 25]$	$[-40, -14]$
Skåne	$[1.5, 5.7]$	$[0.9, 19]$	$[0.4, 2.0]$	$[1.2, 4.1]$	$[-0.7, 11]$	$[-14, -3.7]$
Stockholm	$[-7.0, 4.2]$	$[-9.0, 22]$	$[-1.3, 0.3]$	$[-5.8, 2.2]$	$[-9.5, 2.2]$	$[-4.7, 12]$
Västra Götaland	$[2.1, 7.9]$	$[3.1, 20]$	$[0.2, 2.5]$	$[1.1, 6.7]$	$[-1.1, 17]$	$[-16, -3.4]$

VI. Appendix: Infections and traffic components

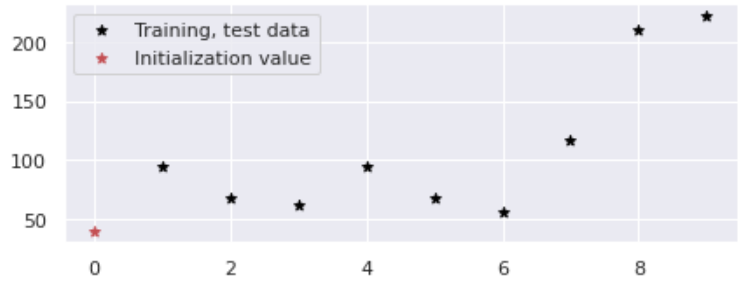

Fig. 1: Pirkanmaa, observed infections

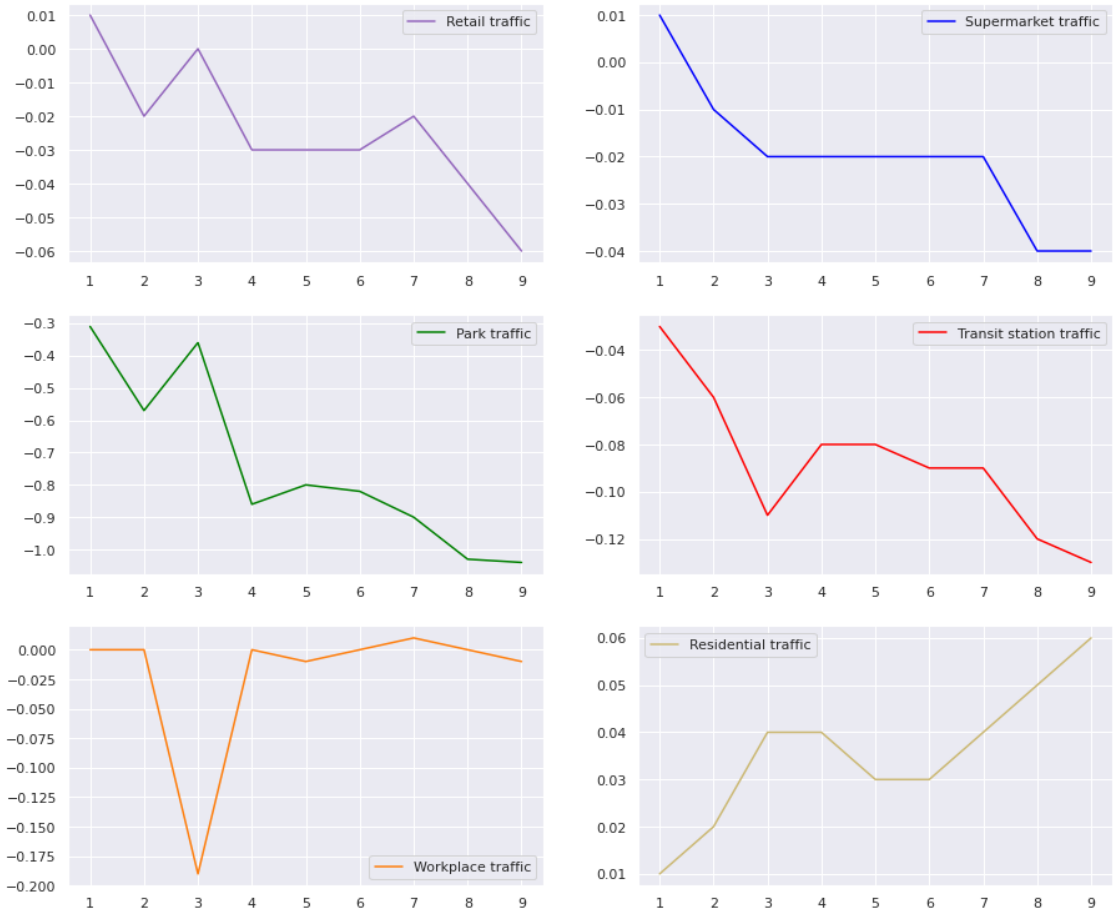

Fig. 2: Pirkanmaa, observed traffic data

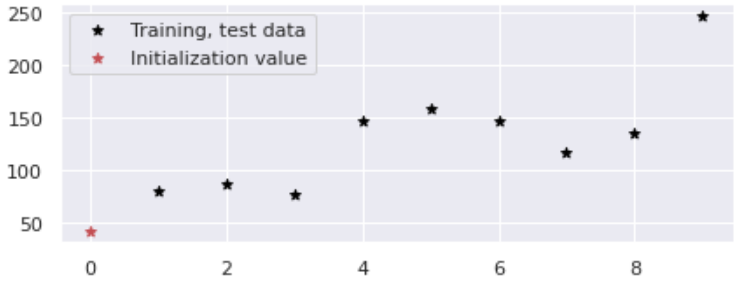

Fig. 3: Southern Finland, observed infections

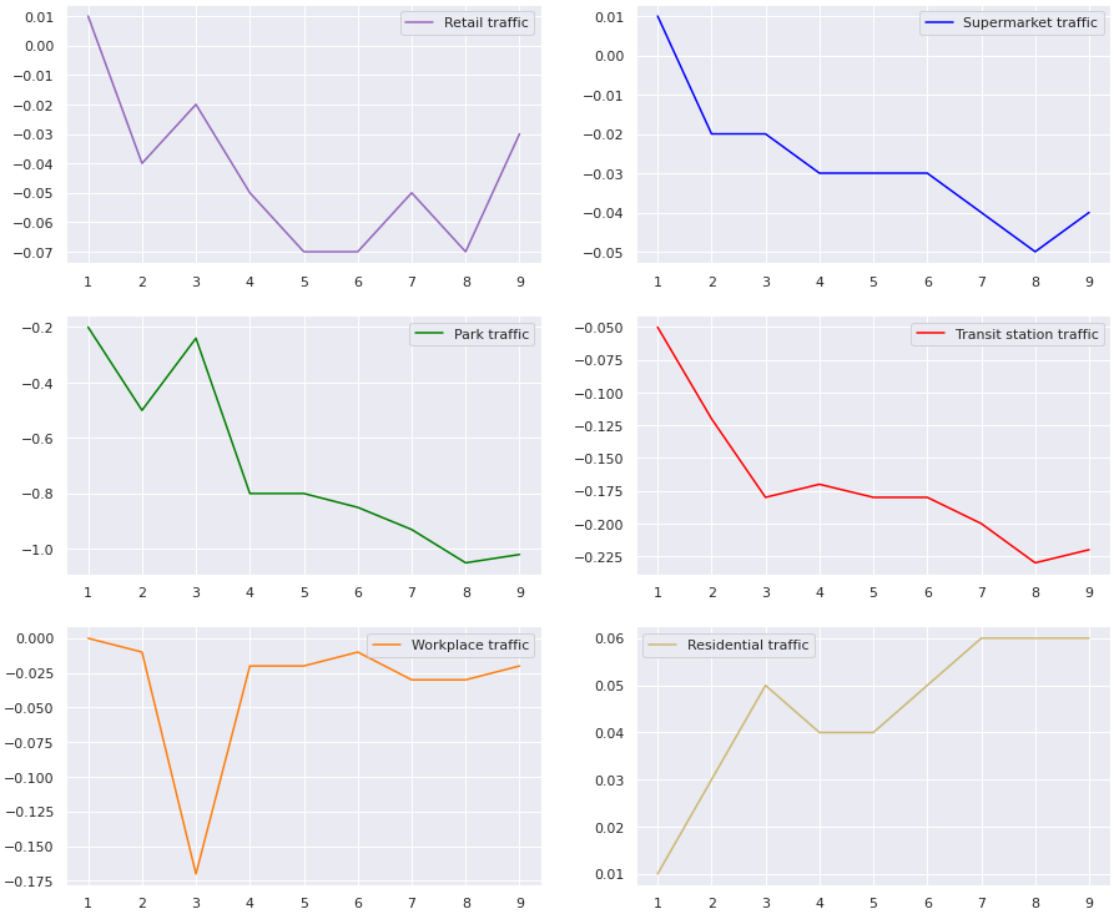

Fig. 4: Southern Finland, observed traffic data

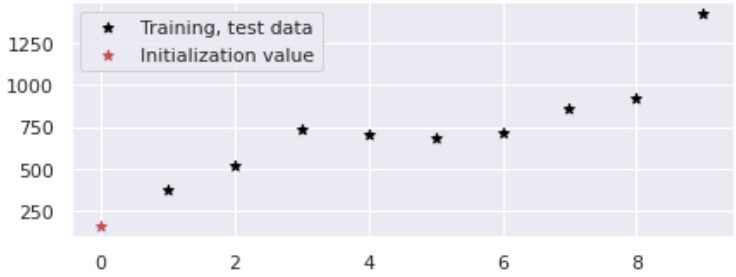

Fig. 5: Uusimaa, observed infections

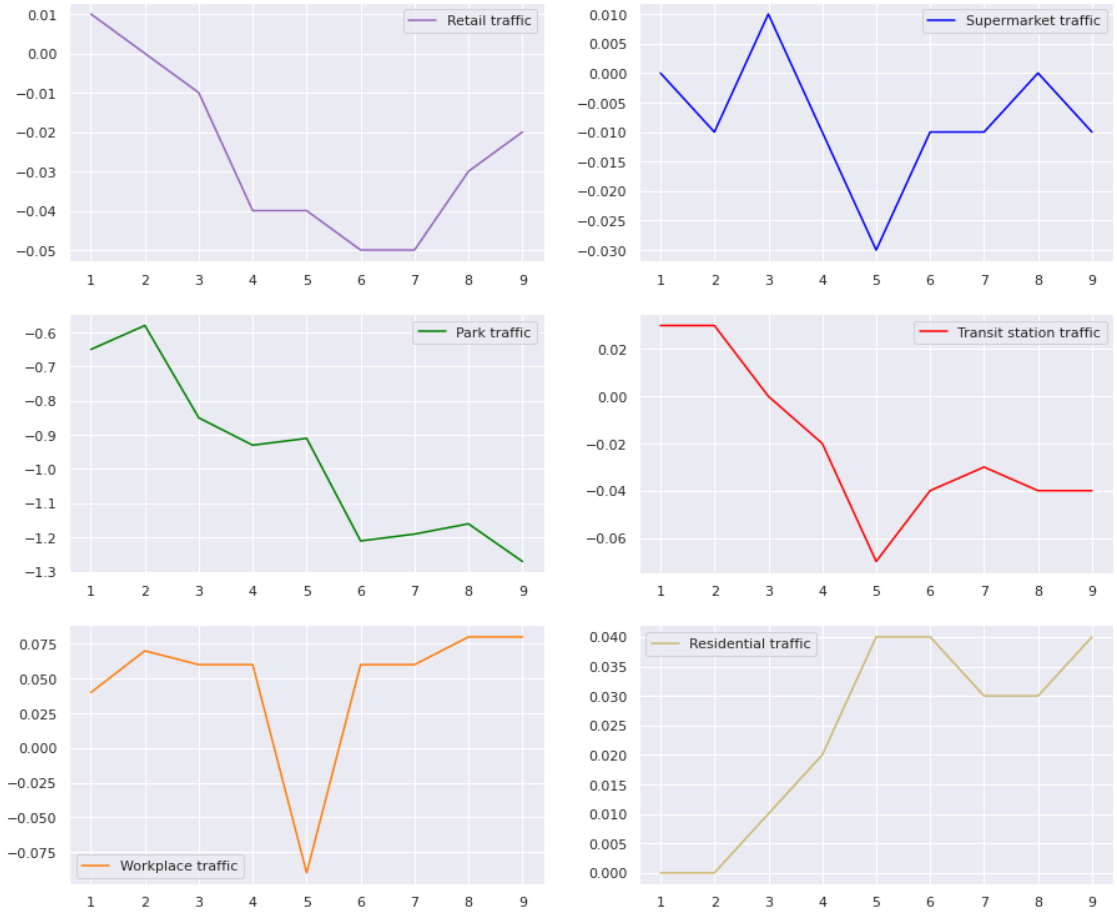

Fig. 6: Uusimaa, observed traffic data

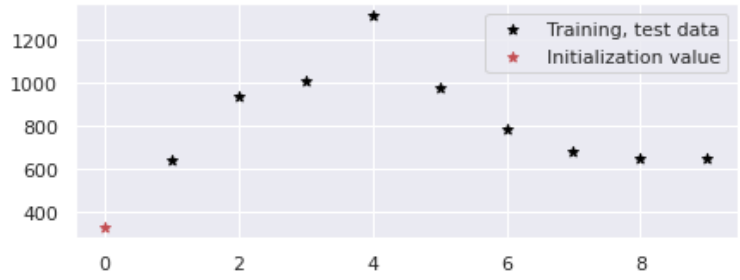

Fig. 7: Oslo county, observed infections

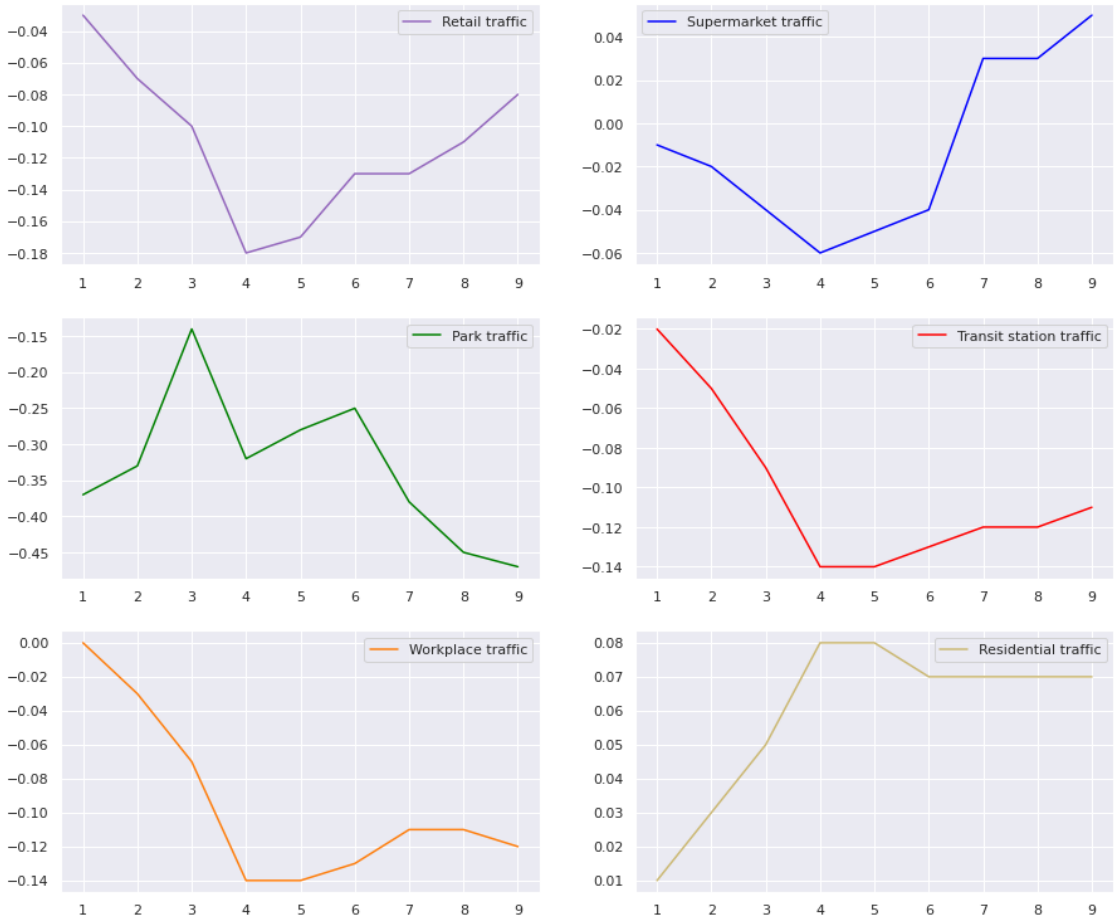

Fig. 8: Oslo county, observed traffic data

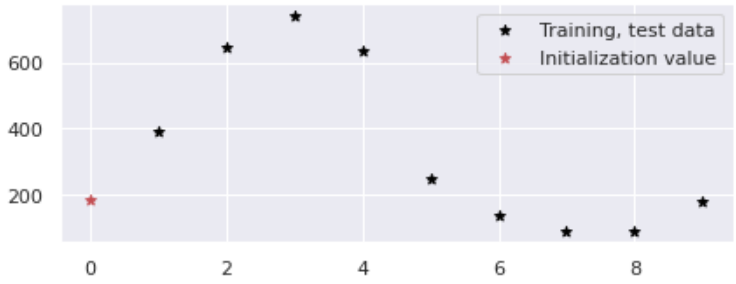

Fig. 9: Vestland, observed infections

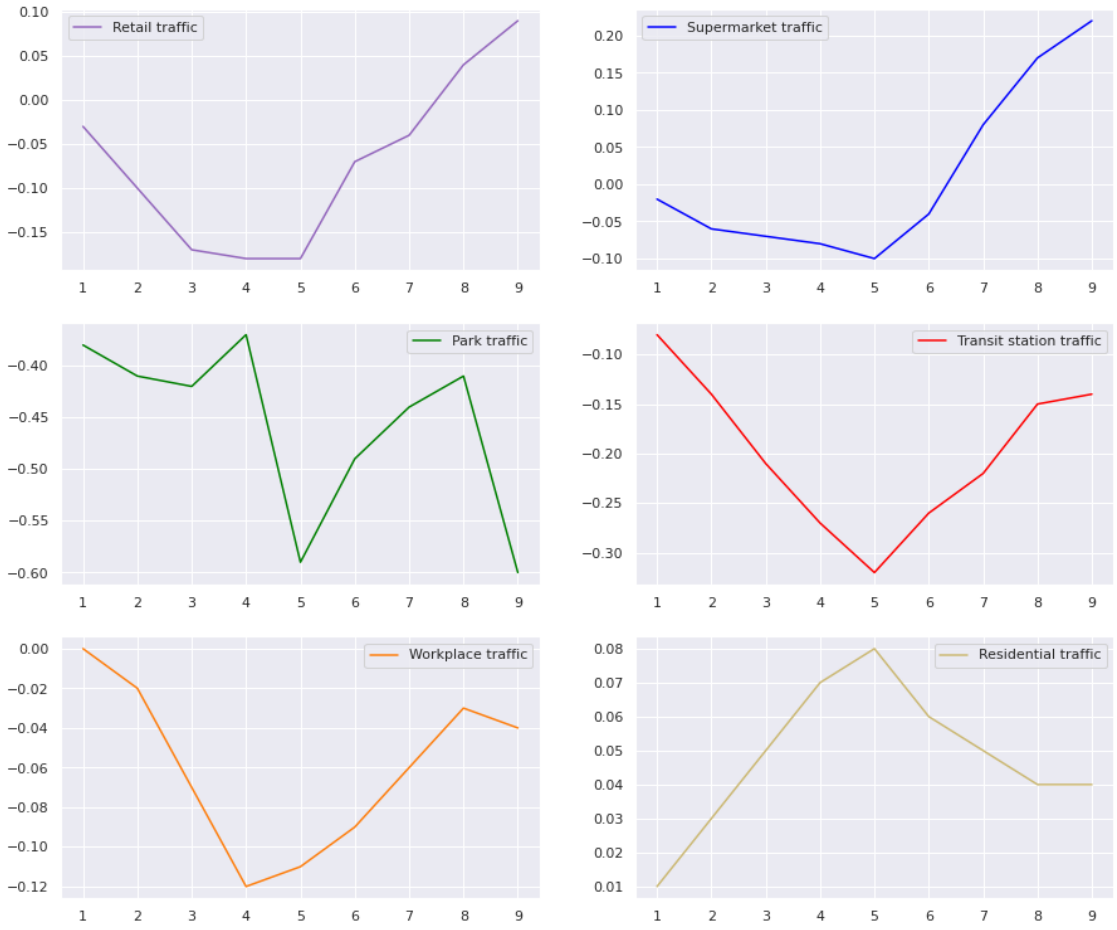

Fig. 10: Vestland, observed traffic data

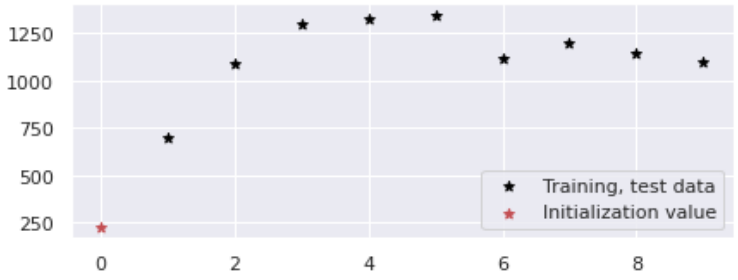

Fig. 11: Viken, observed infections

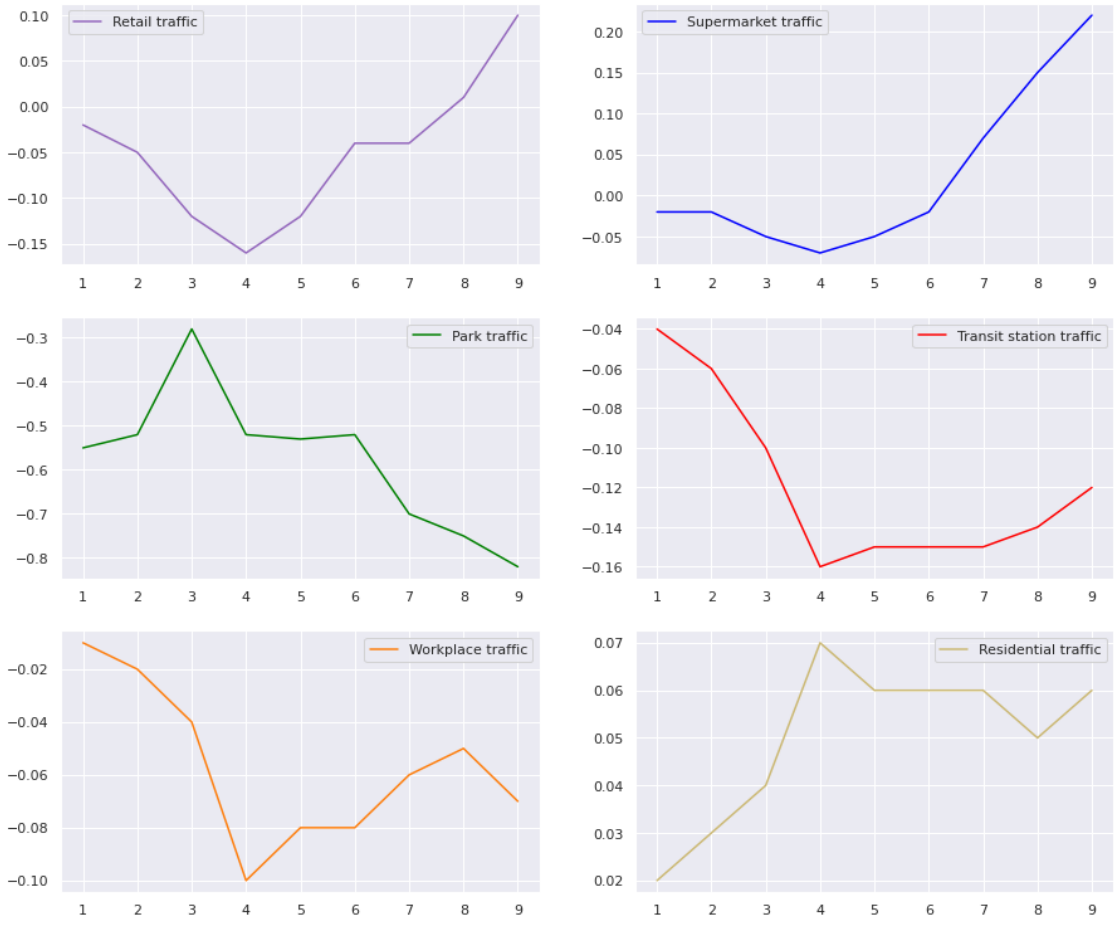

Fig. 12: Viken, observed traffic data

Fig. 13: Skåne, observed infections

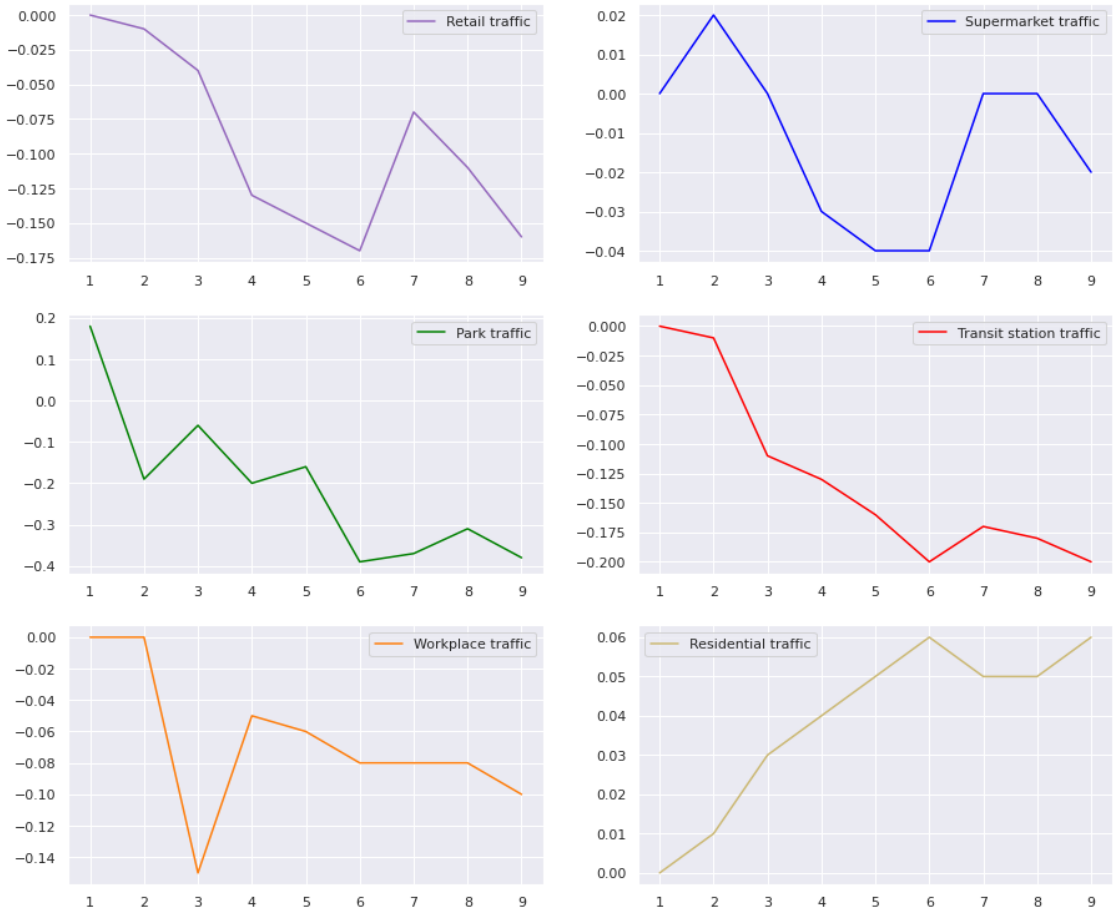

Fig. 14: Skåne, observed traffic data

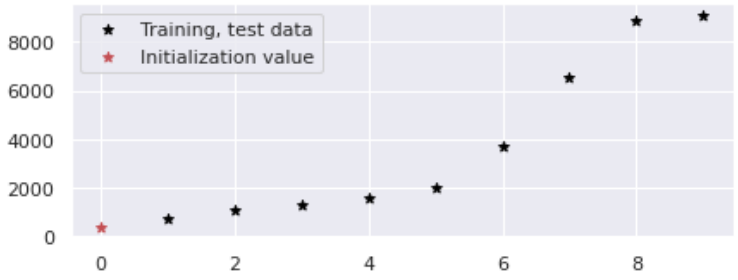

Fig. 15: Stockholm, observed infections

Fig. 16: Stockholm, observed traffic data

Fig. 17: Västra Götaland, observed infections

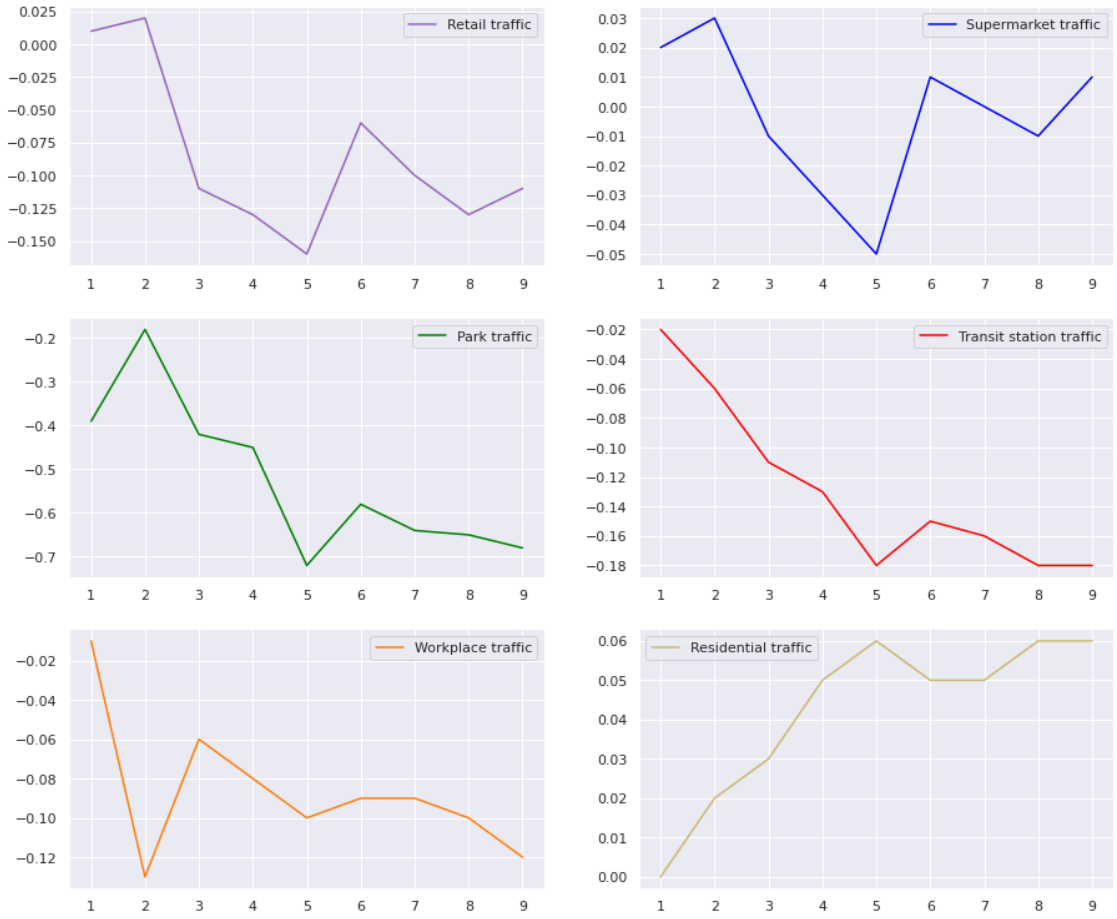

Fig. 18: Västra Götaland, observed traffic data