FEDERAL STATE AUTONOMOUS EDUCATIONAL INSTITUTION

OF HIGHER EDUCATION

ITMO UNIVERSITY

Report on learning practice # 4

Stationarity of the processes

Performed by

Igor Vernyy, j4134c

Kirill Mukhin, j4134c

Alexander Petrov, j4134c

Bogdan Chertkov, j4132c

St. Petersburg

2022

# Substantiation of chosen sampling

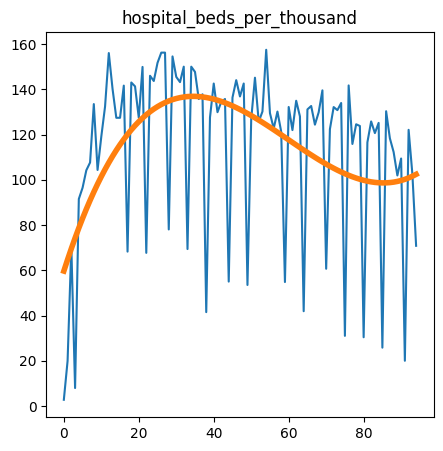
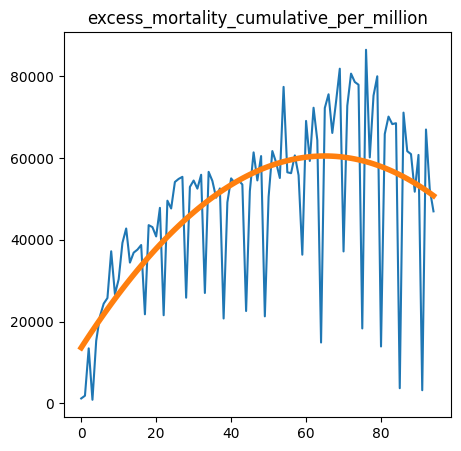
For this lab have chosen 5 variables for the analysis, 3 of them serve as predictors, and 2 - as target variables:

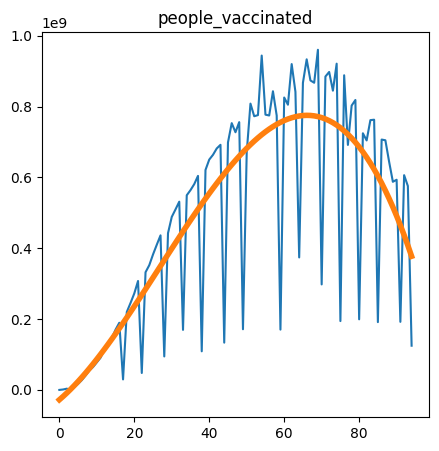
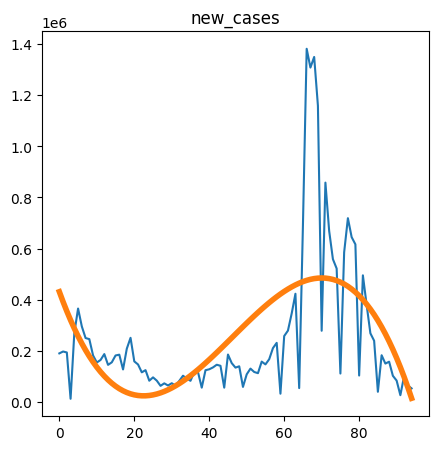
* Predictors:
  + **hospital\_beds\_per\_thousand** - Share of the population with basic hand washing facilities on premises, most recent year available
  + **people\_vaccinated** - Total number of people who received at least one vaccine dose
  + **total\_tests** - New tests for COVID-19 (only calculated for consecutive days)
* Target variables:
  + **excess\_mortality\_cumulative\_per\_million** - Cumulative difference between the reported number of deaths since 1 January 2020 and the projected number of deaths for the same period based on previous years, per million people
  + **new\_cases** - New confirmed cases of COVID-19

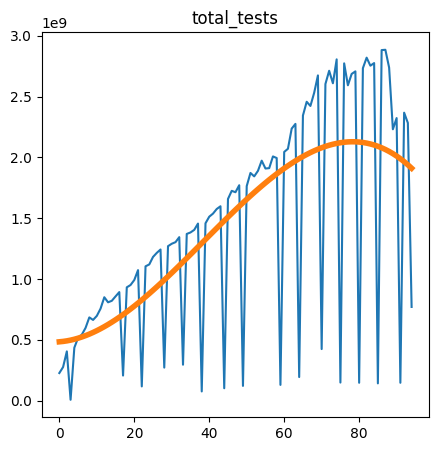
We believe that such factors as the number of tests and vaccinated people along with hospital beds are associated with the number of new cases as well as with the excess mortality

# Stationary analysis

We analyzed stationarity of a process (for mathematical expectation and variance) for all chosen variables. Trend lines for each variable are presented at the graphs below







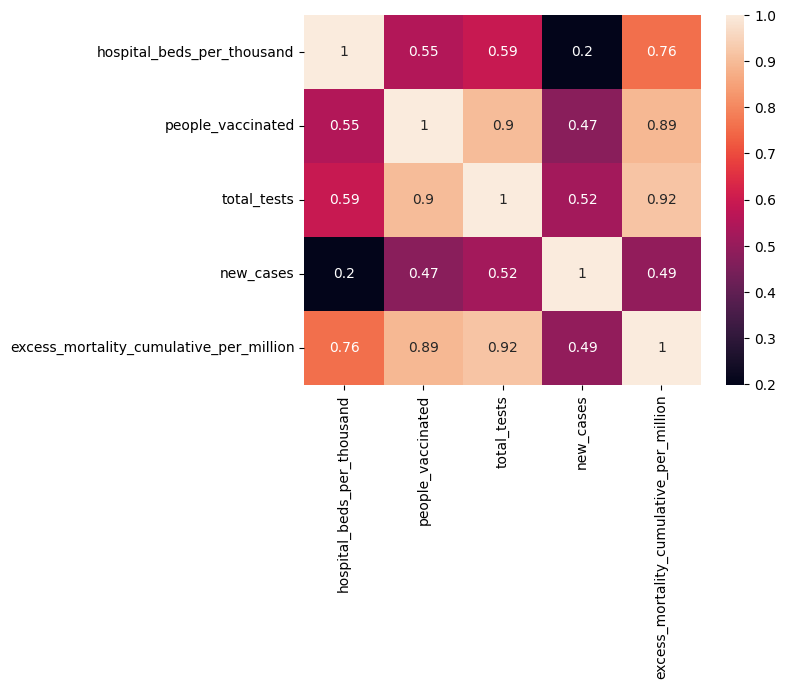
Statistical estimators as mean, standard deviation and variance for each variable are seen at the presented table:

| Variable | Mean | Standard Deviation | Variance |
| --- | --- | --- | --- |
| excess\_mortality\_cumulative\_per\_million | 48 444.53 | 20 996.3 | 440 844 461.65 |
| hospital\_beds\_per\_thousand | 114.04 | 38.08 | 1450.06 |
| new\_cases | 251 353.53 | 279 872.12 | 78 328 401 640.89 |
| people\_vaccinated | 487 096 471.04 | 312 558 553.72 | 9.769284950126229e+16 |
| total\_tests | 1 442 451 407.62 | 897 663 785.67 | 8.058002721014452e+17 |

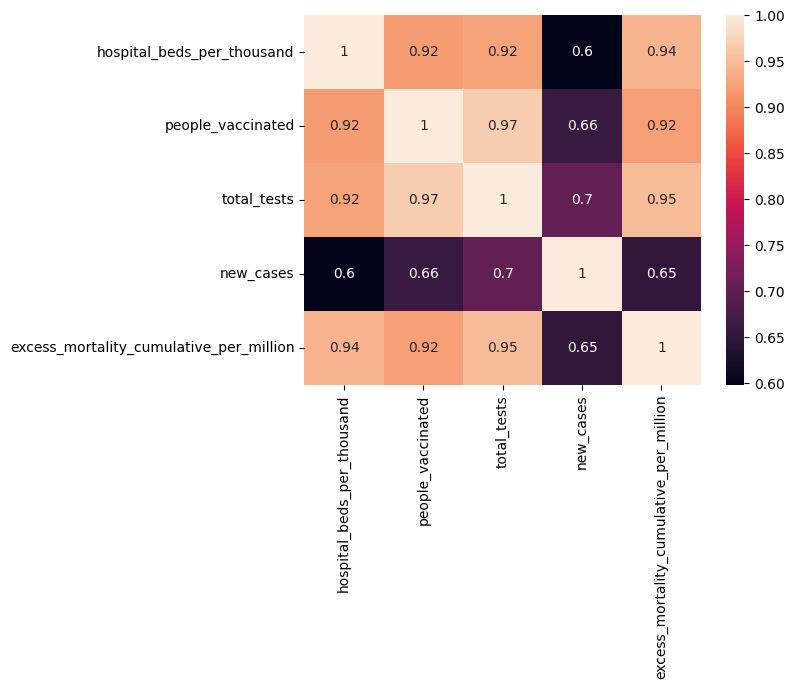
# Covariance or correlation function analysis

We examined the correlation between chosen variables - all of them have statistically significant correlation with p-value < 0.05, the correlation coefficients are all positive and are ranged between 0.2 and 0.92

The lowest correlation (0.2) is between hospital\_beds\_per\_thousand and new\_cases and the highest (0.92) is between excess\_mortality\_cumulative\_per\_million and total\_tests

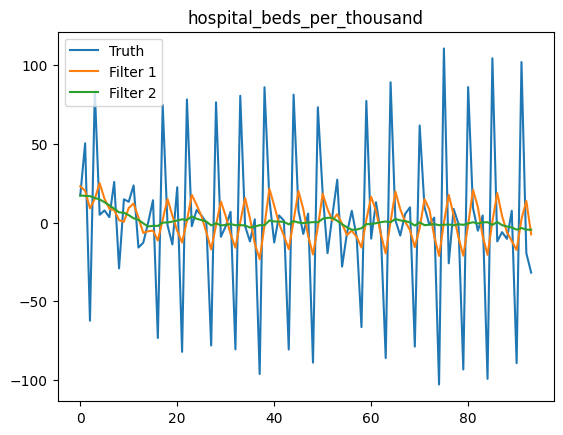
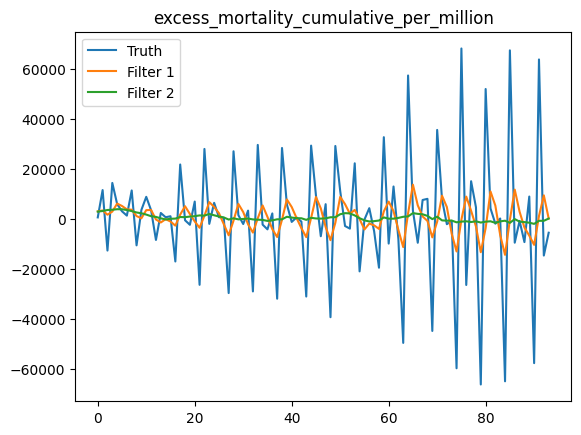


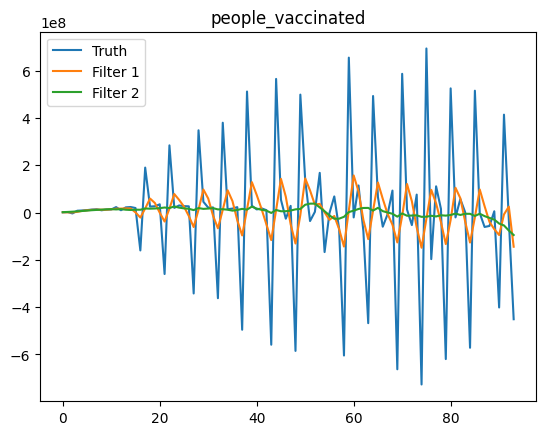
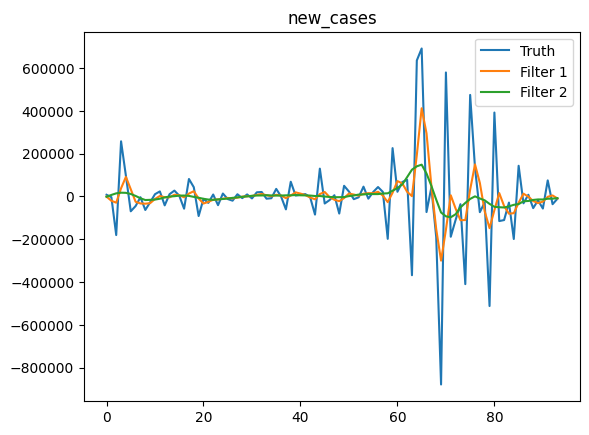
We also examined the correlation between variables on trendless data - all coefficients remained positive, but the range changed to 0.6 (hospital\_beds\_per\_thousand and new\_cases) to 0.97 (people\_vaccinated and total\_tests)

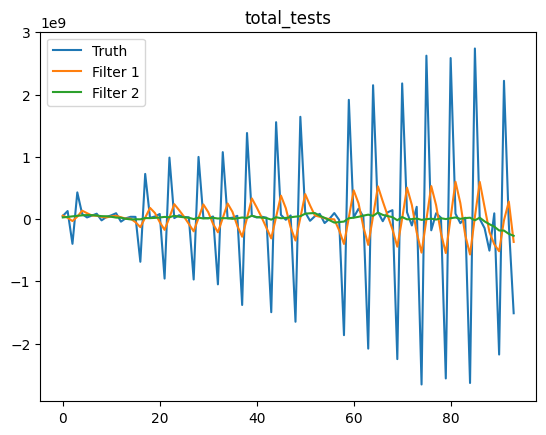


# Noise filtration

We performed 2 filter types for our target variables with windows of 12 and 24 - the resulting distributions are shown on the following graphs

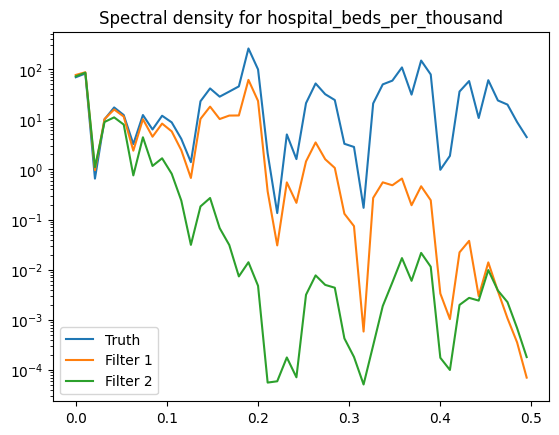
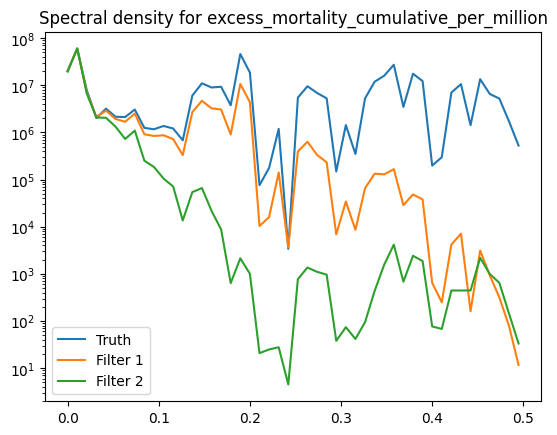


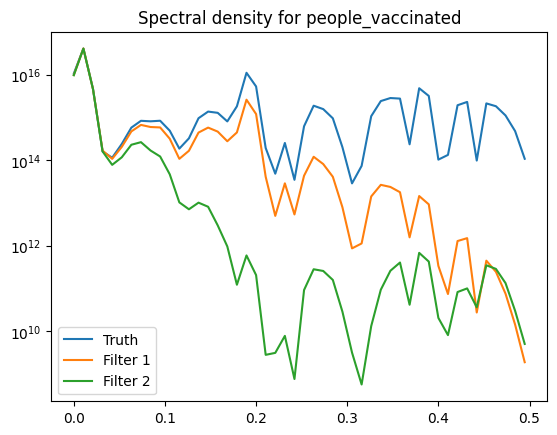
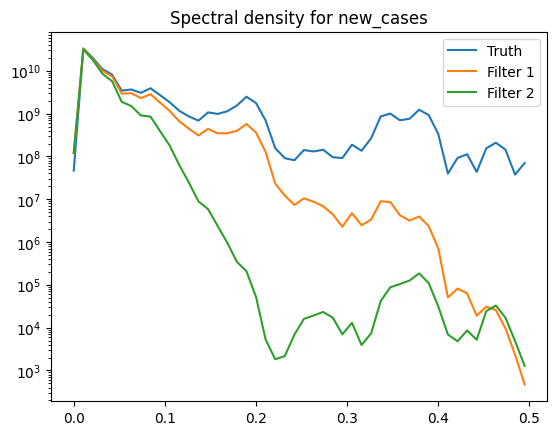


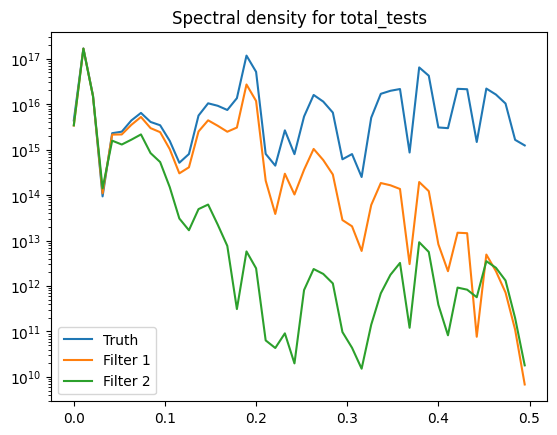


# Estimation of spectral density function

At the next step, we estimated the power spectral density of each variable, graphs of which one can see below. As it is shown in the graphs, after filtering, curves are located under the “Truth” curve, and generally Filter 2 is located below other curves



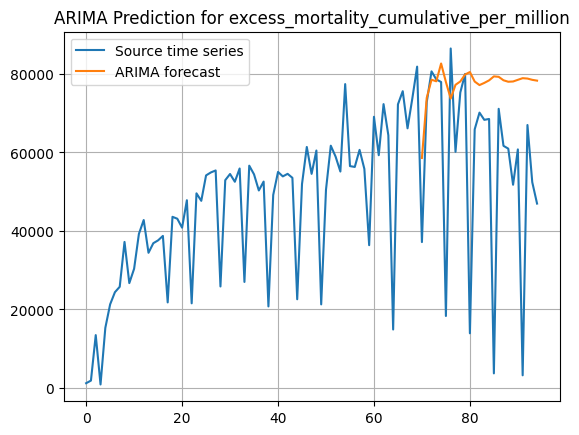
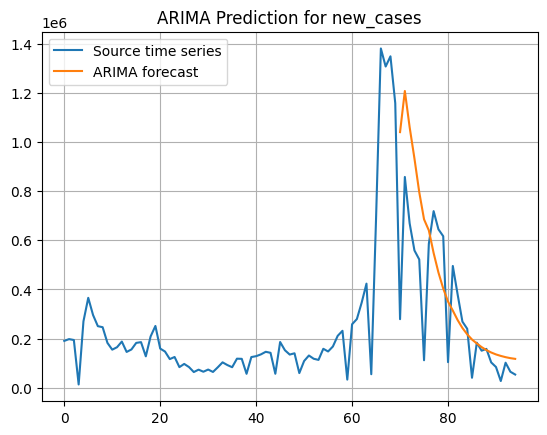




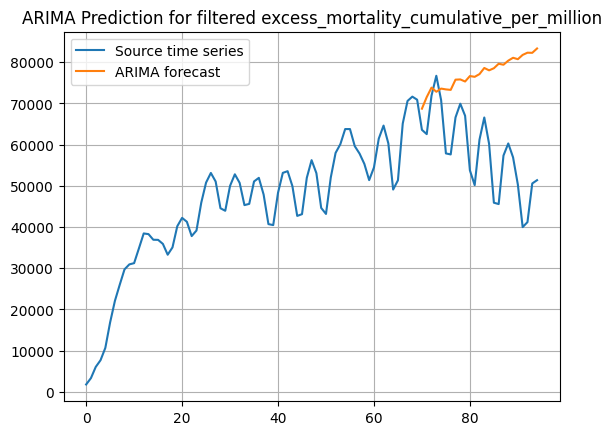
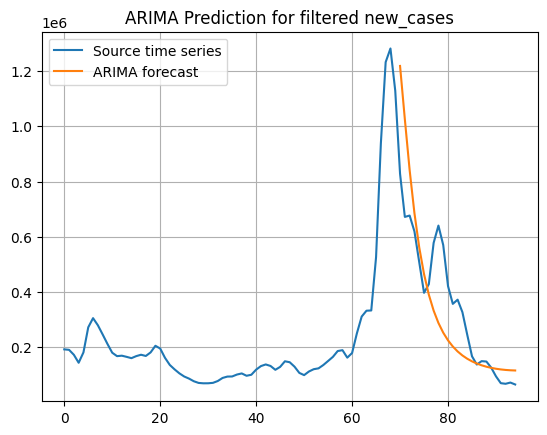
# Auto-regression model

At this point, we built auto-regression model for filtered and non-filtered data. We used the ARIMA model to predict values of target variables. Obtained graphs are presented below. And as one can see from the graph the model presented quite good results for new\_cases target variable, though for excess\_mortality it resulted in less accurate prediction

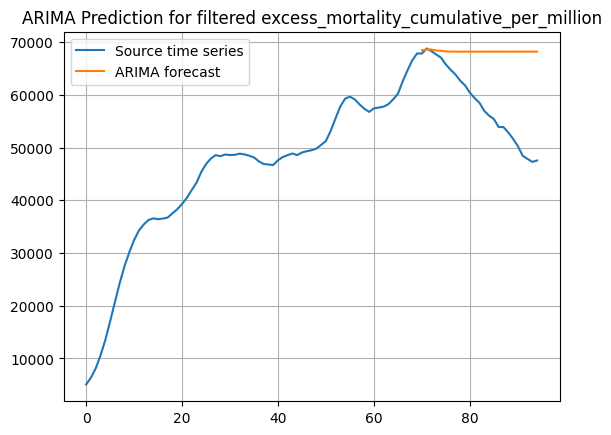
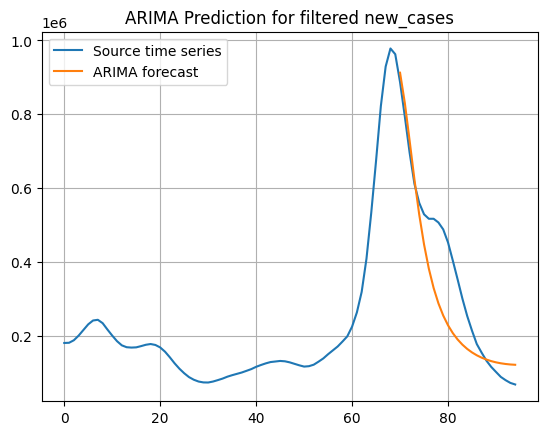
Original Data



Filter 1



Filter 2



| Data Type | Target Variable | R^2 | MSE | MAPE |
| --- | --- | --- | --- | --- |
| Original | new\_cases | -0.044 | 66 430 885 780.92 | 100.806 |
| excess\_mortality\_cumulative\_per\_million | -0.737 | 969 402 201.428 | 227.517 |
| Filter 1 | new\_cases | 0.444 | 29 749 016 070.555 | 35.95 |
| excess\_mortality\_cumulative\_per\_million | -4.585 | 513 373 873.045 | 37.659 |
| Filter 2 | new\_cases | 0.774 | 12 747 671 177.288 | 29.681 |
| excess\_mortality\_cumulative\_per\_million | -1.919 | 147 657 374.35 | 18.783 |

# Model in a form of linear dynamical system

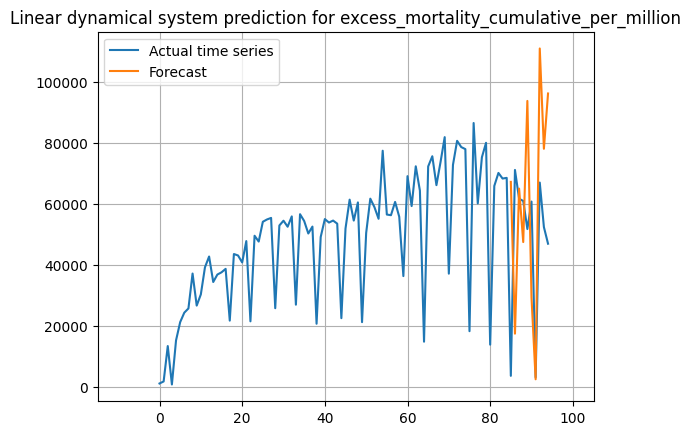
We created two models in a form of linear dynamical systems for both of our target variables - Excess Mortality and New Cases. For both models we used all three of our chosen predictors

Excess Mortality

The pipeline of the linear dynamical system looks like that with three predictors for the target variable:

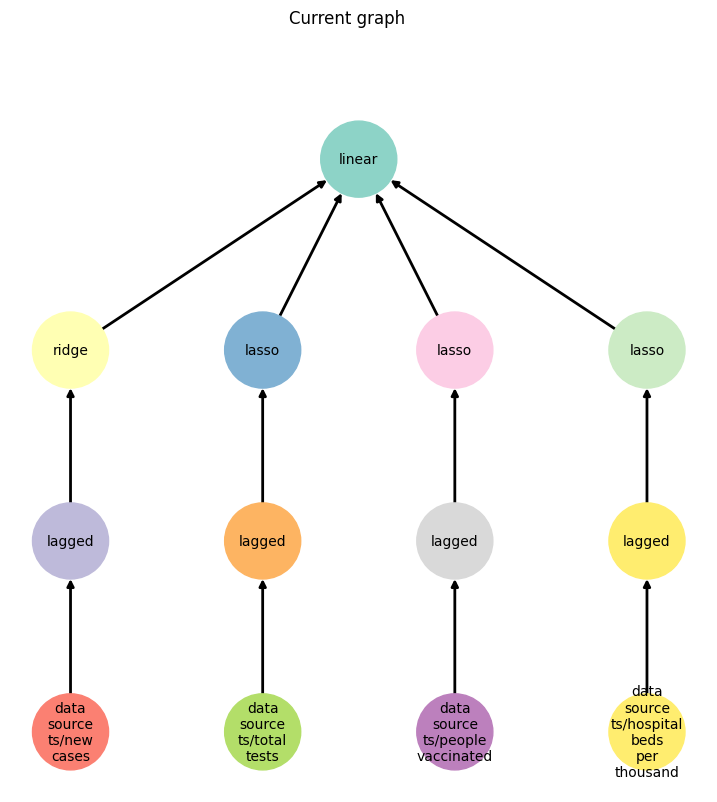


The forecast shows reasonable results as the correlation between predictors and the target variable is high, though the model did not catch the decreasing trend of the last observations

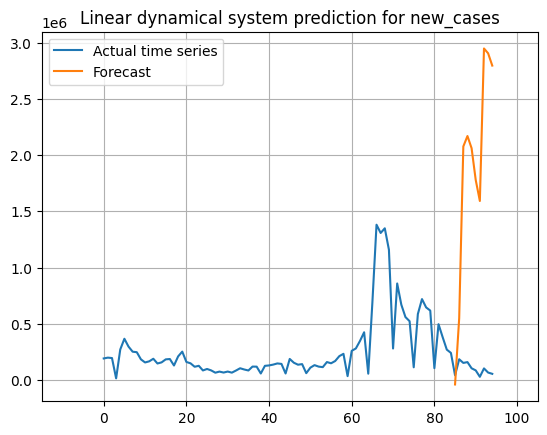


New Cases

Next, we trained the prediction model for another target variable - new\_cases - with the same set of predictors. Just like in the latter case we built the following pipeline



And as one can see from the obtained graph below the quality of the prediction model is extremely low because of the weaker correlation between the target variable and predictors compared to the case where we predicted excess mortality



**Appendix**

GitHub Link: <https://github.com/D3lph1/methods-and-models-for-multivariate-data-analysis/blob/master/Lab%204/lab_4_notebook.ipynb>