FEDERAL STATE AUTONOMOUS EDUCATIONAL INSTITUTION
OF HIGHER EDUCATION
ITMO UNIVERSITY

Report on learning practice #4

Stationarity of the processes

Performed by
Dmitry Grigorev,
Eugenia Khomenko,
Efim Podkovirkin,
Arina Syrchenko

St. Petersburg
2022

Contents

1. Data description

Let D be the modified dataset on Narvik roads. The features here are:

- lat_ — latitude
- lon_ — longitude
- State_ — word description of road state (1: 'dry', 2: 'moist', 3: 'wet', 4: 'icy', 5: 'snowy', 6: 'slushy')
- Ta_mean,Ta_min,Ta_max — atmosphere temperature
- Tsurf_mean,Tsurf_min,Tsurf_max — surface temperature
- Water_mean,Water_min,Water_max — water layerw width (0 – 3 mm)
- Speed_mean,Speed_min,Speed_max — wind speed (in knots, 5 $knots \approx 9.3\ km/h$)
- Height_mean,Height_min,Height_max — height of location above mean sea level
- Tdew_mean,Tdew_min,Tdew_max — dew point ($Celsius$)
- Friction_mean,Friction_min,Friction_max — friction value (0 – 1, 0 means no friction)
- Date,Time, date_time, FullDate — time and date
- Direction_min,Direction_max — wind direction ($degrees$)
- ClosestCity, location
- maxtempC,mintempC — day maximum and minimum of temperature ($Celsius$)
- totalSnow_cm — total snowfall (cm)
- sunHour — passed sun energy in $Sun-Hours$ (A $Sun-Hour$ is "1000 watts of energy shining on 1 square meter of surface for 1 hour")
- uvIndex — ultraviolet index
- moon_illumination — moon phase ($percents$)
- moonrise — time of Moon rise
- moonset — time of Moon set

- sunrise — time of Sun rise
- sunset — time of Sun set
- DewPointC — hourly dew point measurement (*Celsius*)
- FeelsLikeC — hourly Feels-like temperature (*Celsius*)
- HeatIndexC — hourly heat index (*Celsius*)
- WindChillC — hourly wind-chill index (*Celcius*)
- WindGustKmph — hourly wind gust measure (km/h)
- cloudcover — hourly cloud cover index (*percents*)
- humidity — hourly humidity (*percents*)
- precipMM — hourly precipitation (mm)
- pressure — hourly atmosphere pressure ($mbar$)
- tempC — hourly atmosphere temperature (*Celsius*)
- visibility — hourly visibility (0–10, 0 means poor visibility)
- winddirDegree — hourly wind direction (*degrees*)
- windspeedKmph — hourly wind speed (km/h)

2. Substantiation of chosen sample

Friction_mean, Water_mean and Tsurf_mean are chosen as targets. As predictors we chose variables Height_mean and Speed_mean. The data are collected across the map in fig. 1 and they are too inhomogeneous and non-equidistant in time.

To tackle the latter problem, we did downsample according to the time with the period of 3 minutes with mean aggregation of the variables. This operation generated some missing data which occurred due to the time intervals between two consequential observation larger than 3 minutes in time. Firstly, we selected only first 300 observations from the beginning of the measurements to study only the corresponding geographic location (neighborhood of Øyjord town). Secondly, there are a few of missing data which we filled by rolling mean of order 3 since the gaps are small and the close observations are likely to be similar.

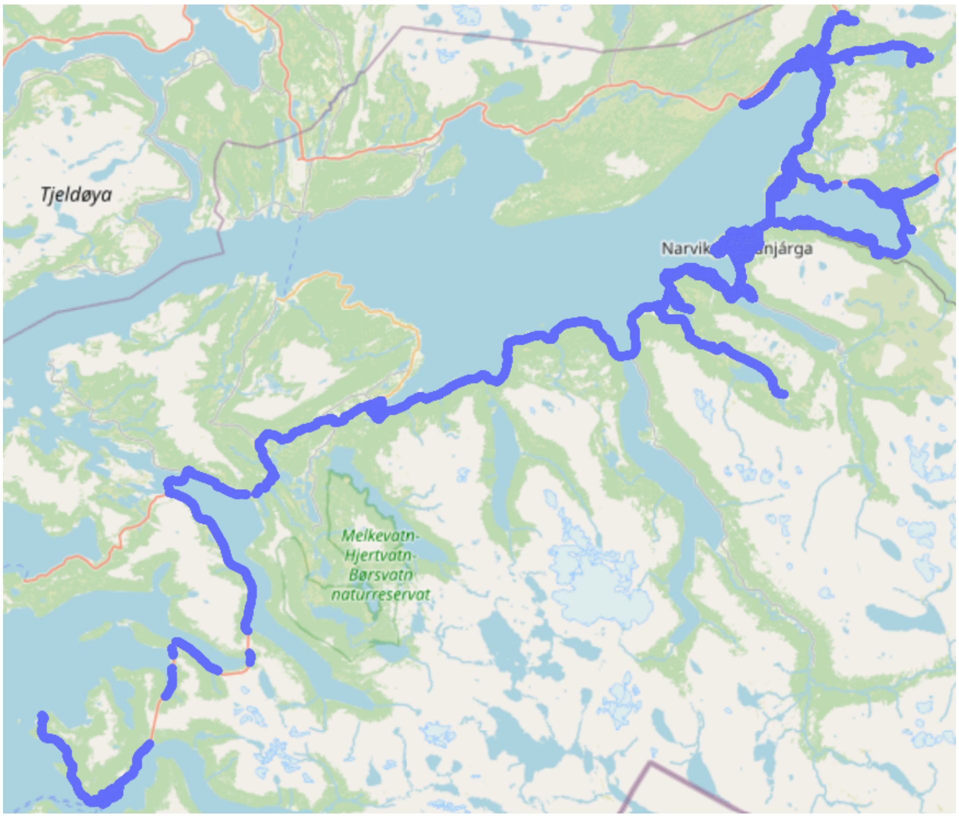

Figure 1: Geography of the data

3. Stationary and covariance/correlation function analysis

At first, we drew the series themselves with the fitted polynomials of the 9th degree which approximate the series' possible trends. They are all shown in figure 2.

Although Friction_mean serial looks stationary, we intentionally removed both its learnt trend and the trends of other features each of which has a certain tendency. The resulted series are shown in figure 3 where for each of the variable under the study the estimates of mean and variance functions of the series are provided to infer weak stationarity.

One can observe that the estimates of mean functions looks constant for all of the modified series. As for the variance, as soon as it stabilizes, it fluctuates a bit near some horizontal line for all of the features. Also we analyzed autocovariance functions for the features which are presented in figure 4. All of them look approximately constant what also shows that the processes are stationary in the weak sense. To conclude the topic of stationarity, we did Augmented Dickey-Fuller (ADF) to finalize our conclusion.

ADF test checks the hypothesis on whether the given time series is I(1) (i.e. the time series of its differences is stationary but it is not) versus I(0) alternative hypothesis assuming that its structure is described by an ARMA model white noise errors. The test estimates the regression coefficient ϕ

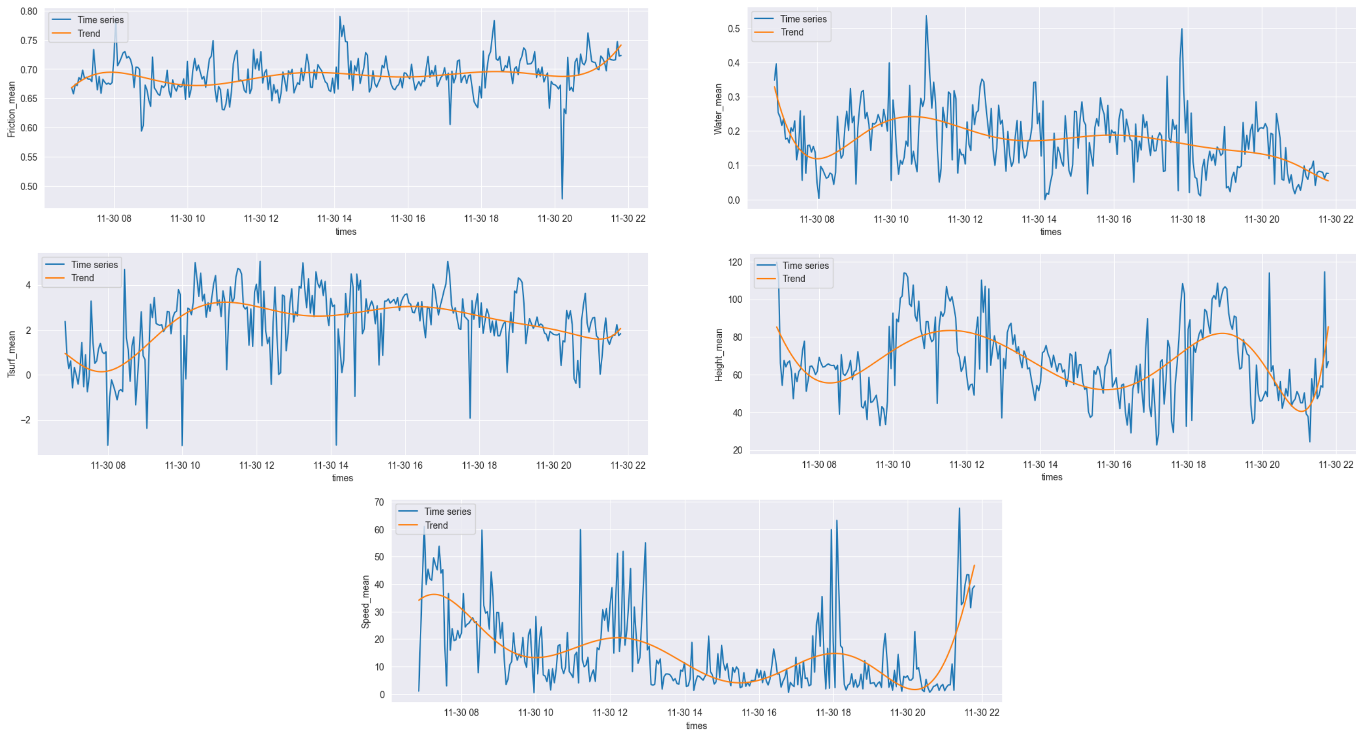

Figure 2: The histogram of the generated sample vs the mixture's density and the histogram of observed Friction_mean

and the regression:

$$y_t = \beta^T \mathbf{D}_t + \phi y_{t-1} + \sum_{j=1}^{q} \psi_j \Delta y_{t-j} + \varepsilon_t,$$

where β is a vector of coefficients for components in $\mathbf{D}_t$ which determines deterministic part of the series, i.e. trend or seasonality, $\Delta y_\tau = y_\tau - y_{\tau-1}$. Given the OLS estimates $\widehat{\phi}$ and $\widehat{\psi}_k$ of the coefficients ϕ and ψ_k, the test statistic is

$$t_{\text{ADF}} = T \frac{\widehat{\phi} - 1}{1 - \sum_{j=1}^{q} \widehat{\psi}_j},$$

where T is the time series' time range. The test statistic distribution has its own tabulated distribution and limiting distribution under some assumptions. The parameter q determines the lag length in the aforementioned regression. If it is too large, the test's power suffers. If it is too small, the test becomes biased [1].

The results of the test for our time series are as follows:

- ADF test rejects the hypothesis on Friction_mean having unit root, p-value $\approx 3.2 \cdot 10^{-11}$ with $q = 1$;
- the same for Water_mean, p-value $\approx 3.24 \cdot 10^{-9}$ with $q = 5$;
- the same for Tsurf_mean, p-value $\approx 5.8 \cdot 10^{-25}$ with $q = 0$;

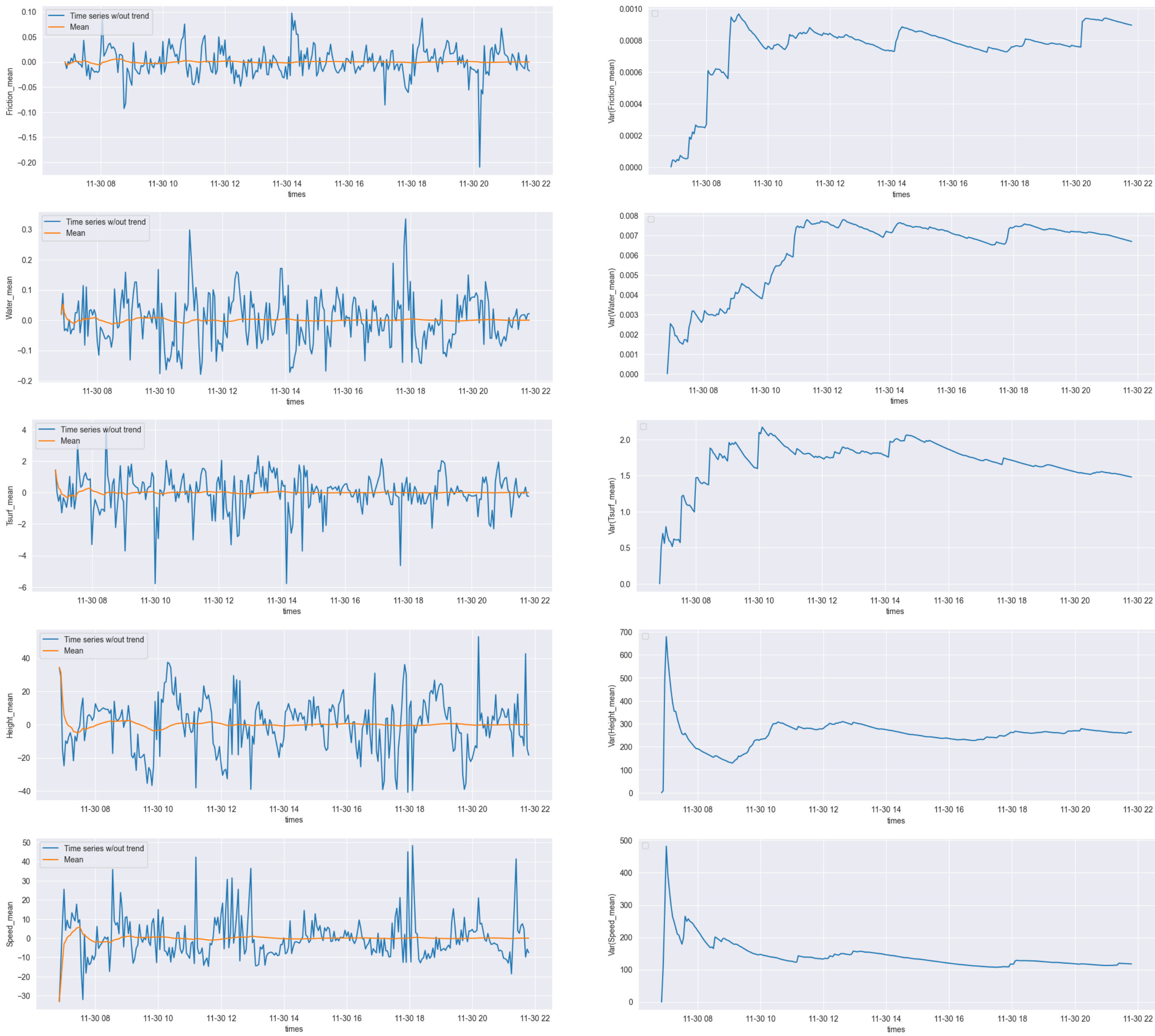

Figure 3: The histogram of the generated sample vs the mixture's density and the histogram of observed Friction_mean

- the same for Height_mean, p-value $\approx 1.0 \cdot 10^{-6}$ with $q = 4$;
- the same for Water_mean, p-value $\approx 1.9 \cdot 10^{-9}$ with $q = 1$.

So we are allowed to treat the resulted time series as stationary ones.

4. Cross-correlation function analysis

We did the research of cross-correlation between our targets and predictors. Their plots are presented in figure 5 and they show no strong linear connections between the variables.

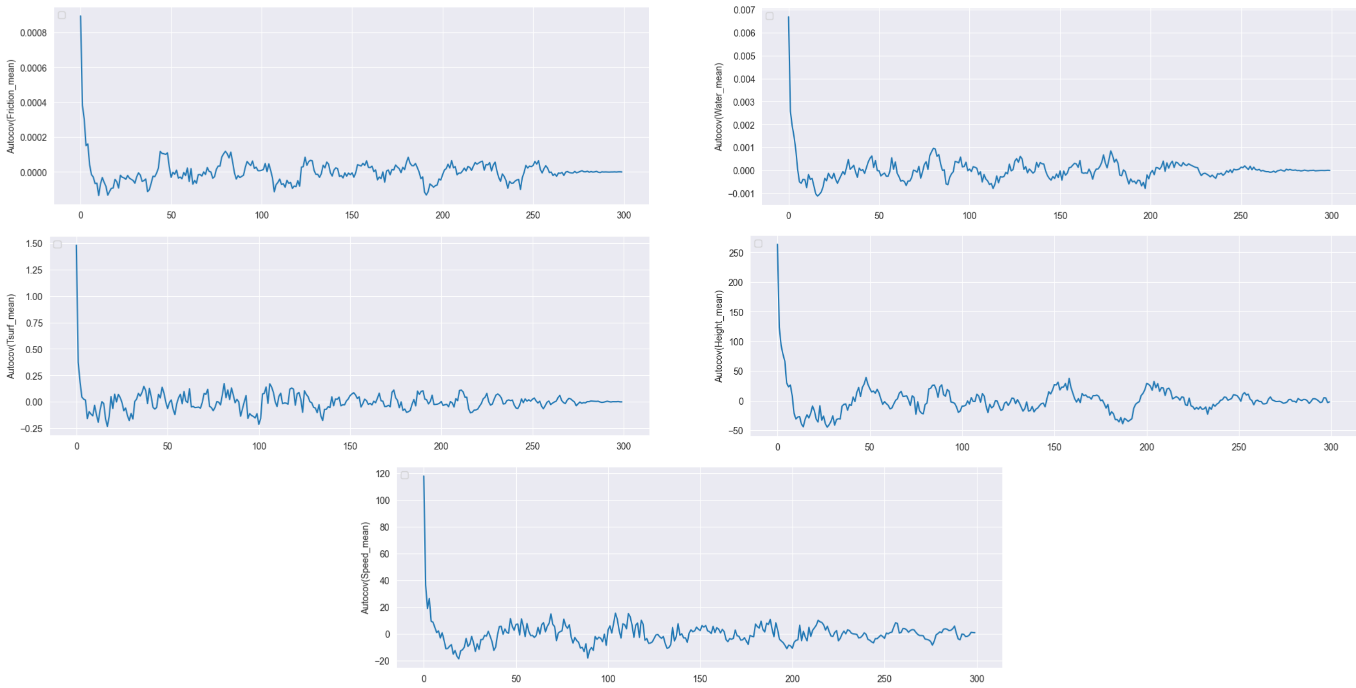

Figure 4: The histogram of the generated sample vs the mixture's density and the histogram of observed Friction_mean

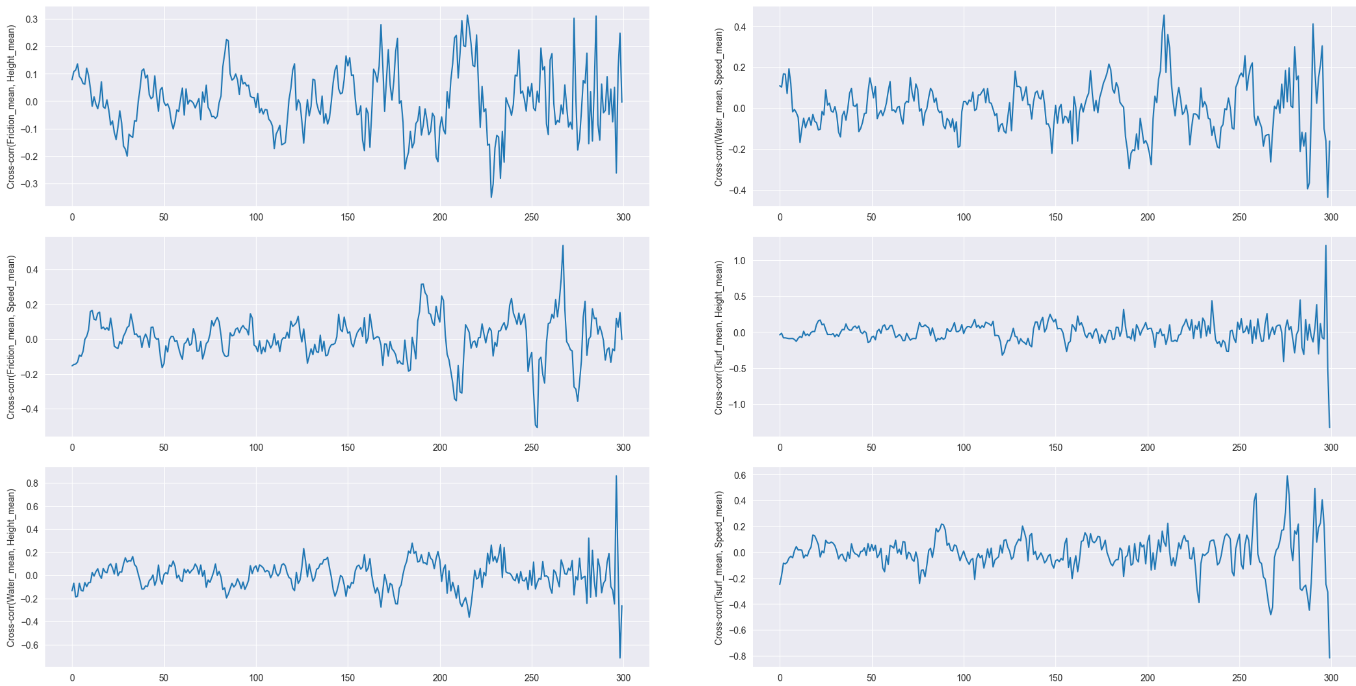

Figure 5: The histogram of the generated sample vs the mixture's density and the histogram of observed Friction_mean

5. Noise filtration

6. Estimation of spectral density function

7. Auto-regression model

8. Model in a form of linear dynamical system

9. Appendix

The Python notebook related to the aforementioned calculations is presented in Github [2].

Bibliography

1. Zivot E., Wang J. Modeling Financial Time Series with S-Plus. — 2001. — P. 120–121. — ISBN: 978-0-387-27965-7.
2. Grigorev D. Code repository. — https://github.com/dmitry-grigorev/MultivarAnalysis/blob/master/Lab4/Lab4notebook.ipynb. — 2022.