Venice is drowning

a Streaming Data Management and Time Series Analysis project

DARIO BERTAZIOLI
FABRIZIO D'INTINOSANTE
UNIVERSITÀ DEGLI STUDI DI MILANO BICOCCA

INTRODUCTION

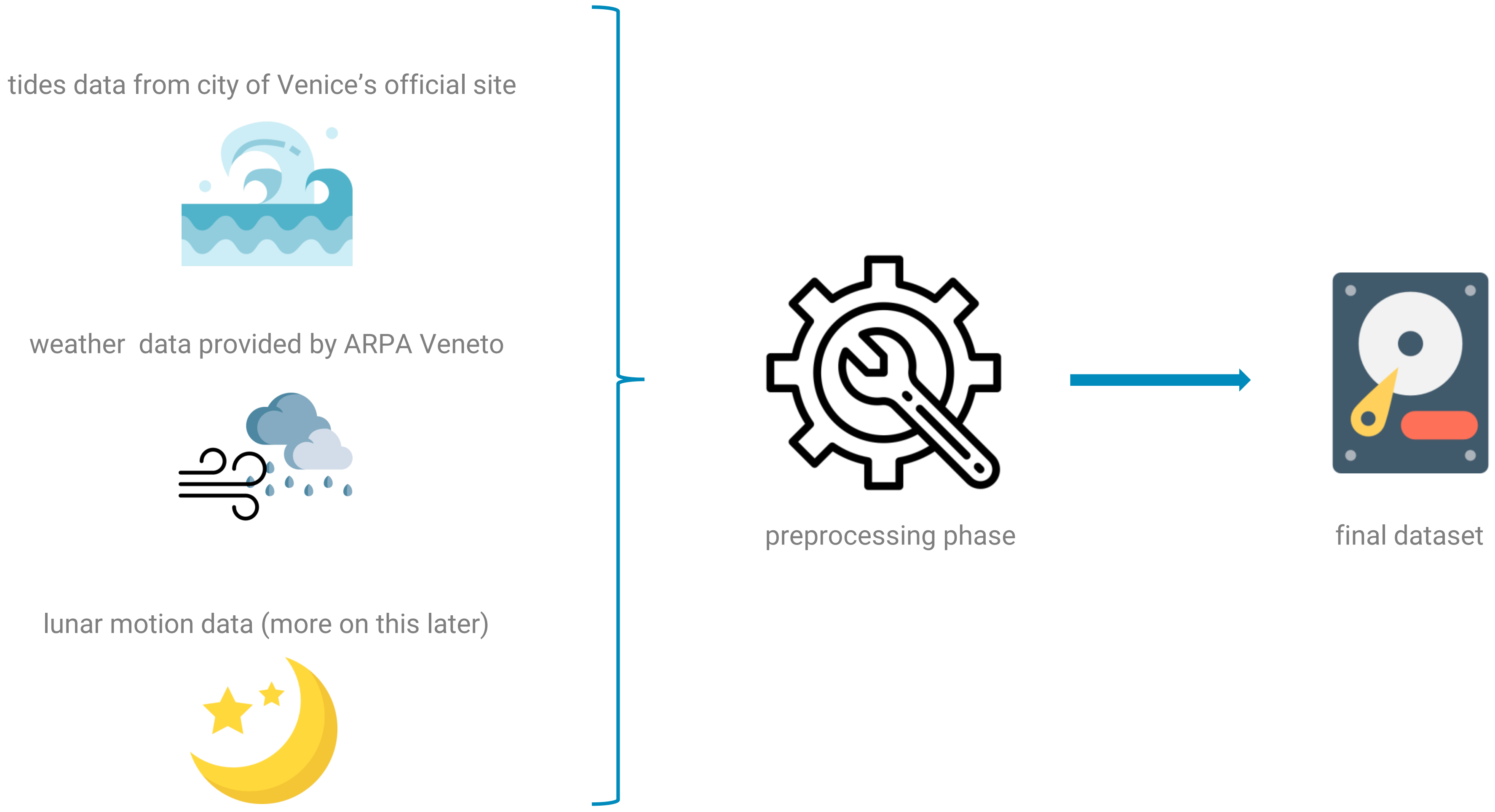
monitoring **tidal levels** is a fundamental task for a city like Venice

through the analysis of the historical series it is possible to try to predict **anomalous peaks** in the high tide level in order to better prepare the city

our objective for this project is to analyze the data of the tide detections regarding the area of the Venice lagoon in order to realize **predictive models**

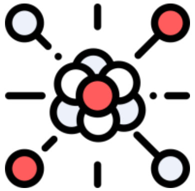
DATA

DATA SOURCES



IN DETAILS

data about tides are available from 1983 to 2018 (soon also 2019)



every year is available into a single file so this required parsing operation

the data are provided in hourly observations and represents the centimeters of sea level compared to a sensor so their range is between [-50, +160]

IN DETAILS

the weather data are provided, on request, by ARPA Veneto

the file provided in particular contains hourly observations about:

- **rain volume** in mm
- **wind direction** in grades
- **wind speed** in m/s

the reference period is 2000-2019 and contains a certain quantity of missing values

multiple imputation using addictive regression, bootstrapping and predictive mean matching (exploiting also tides data to improve conditional imputation)

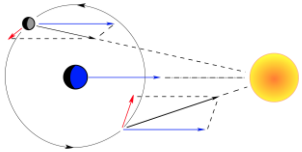


IN DETAILS

core idea: tidal phenomena influenced by the lunar motion (gravitational attraction ~ $1/r^2$)

variation of the Moon distance from Venice could explain part of the time series?

theoretical approach: three body problem
gravitational interactions among the Moon, the Earth and the Sun



"Newtonian Dynamics" by Richard Fitzpatrick



IN DETAILS

Analytical solution:

coordinate change (from Sun-centered to Warth-centered cs)

$$\begin{aligned} \ddot{\mathbf{r}} &= -n^2 a^3 \frac{\mathbf{r}}{|\mathbf{r}|^3} + n'^2 a'^3 \left[\frac{(\mathbf{r}' - \mathbf{r})}{|\mathbf{r}' - \mathbf{r}|^3} - \frac{\mathbf{r}'}{|\mathbf{r}'|^3} \right], \\ \ddot{\mathbf{r}}' &= -n'^2 a'^3 \frac{\mathbf{r}'}{|\mathbf{r}'|^3}, \end{aligned}$$

$$\begin{aligned} \mathbf{r} &= \mathbf{r}_M - \mathbf{r}_E, \\ \mathbf{r}' &= -\mathbf{r}_E, \end{aligned}$$

rotating system (Earth-centered)

$$\ddot{\mathbf{r}} + 2\,\boldsymbol{\omega} \times \dot{\mathbf{r}} + \boldsymbol{\omega} \times (\boldsymbol{\omega} \times \mathbf{r}) = -n^2 a^3 \frac{\mathbf{r}}{|\mathbf{r}|^3} + n'^2 a'^3 \left[\frac{(\mathbf{r}' - \mathbf{r})}{|\mathbf{r}' - \mathbf{r}|^3} - \frac{\mathbf{r}'}{|\mathbf{r}'|^3} \right],$$

expansion in $a/a' = 0.00257$

$$\ddot{\mathbf{r}} + 2\,\boldsymbol{\omega} \times \dot{\mathbf{r}} + \boldsymbol{\omega} \times (\boldsymbol{\omega} \times \mathbf{r}) \simeq -n^2 a^3 \frac{\mathbf{r}}{|\mathbf{r}|^3} + \frac{n'^2 a'^3}{|\mathbf{r}'|^3} \left[\frac{(3\,\mathbf{r} \cdot \mathbf{r}')\,\mathbf{r}'}{|\mathbf{r}'|^2} - \mathbf{r} \right].$$

in Cartesian Coordinates

$$\begin{aligned} X &= x_1/a, \\ Y &= y_1/a, \\ Z &= z_1/a, \\ m &= n'/n = 0.07480 \end{aligned}$$

$$\begin{aligned} \ddot{X} - 2\,\dot{Y} - (1 + m^2/2)\,X &\simeq -\frac{X}{R^3} + \frac{3}{2}\,m^2\,\cos[2\,(1-m)\,T]\,X \\ &\quad -\frac{3}{2}\,m^2\,\sin[2\,(1-m)\,T]\,Y, \\ \ddot{Y} + 2\,\dot{X} - (1 + m^2/2)\,Y &\simeq -\frac{Y}{R^3} - \frac{3}{2}\,m^2\,\sin[2\,(1-m)\,T]\,X \\ &\quad -\frac{3}{2}\,m^2\,\cos[2\,(1-m)\,T]\,Y, \\ \ddot{Z} + m^2\,Z &\simeq -\frac{Z}{R^3}, \end{aligned}$$

IN DETAILS

Perturbation (second order expansion in delta)

$$
\begin{aligned}
\delta\ddot{X} - 2\,\delta\dot{Y} - 3\,(1+m^2/2)\,\delta X &\simeq \frac{3}{2}\,m^2\,\cos[2\,(1-m)\,T] + \frac{3}{2}\,m^2\,\cos[2\,(1-m)\,T]\,\delta X \\
&\quad - \frac{3}{2}\,m^2\,\sin[2\,(1-m)\,T]\,\delta Y - 3\,\delta X^2 + \frac{3}{2}\,(\delta Y^2 + \delta Z^2),
\end{aligned}
$$

$$
\begin{aligned}
\delta\ddot{Y} + 2\,\delta\dot{X} &\simeq -\frac{3}{2}\,m^2\,\sin[2\,(1-m)\,T] - \frac{3}{2}\,m^2\,\sin[2\,(1-m)\,T]\,\delta X \\
&\quad - \frac{3}{2}\,m^2\,\cos[2\,(1-m)\,T]\,\delta Y + 3\,\delta X\,\delta Y, \\
\delta\ddot{Z} + (1+3\,m^2/2)\,\delta Z &\simeq 3\,\delta X\,\delta Z.
\end{aligned}
$$

$$
\begin{aligned}
X &= X_0 + \delta X, \\
Y &= \delta Y, \\
Z &= \delta Z,
\end{aligned}
$$

$$X_0 = (1+m^2/2)^{-1/3}$$

$$|\delta X|,\ |\delta Y|,\ |\delta Z| \ll X_0.$$

$$
\begin{aligned}
\delta\ddot{X} - 2\,\delta\dot{Y} - 3\,(1+m^2/2)\,\delta X &\simeq R_X, \\
\delta\ddot{Y} + 2\,\delta\dot{X} &\simeq R_Y, \\
\delta\ddot{Z} + (1+3\,m^2/2)\,\delta Z &\simeq R_Z,
\end{aligned}
$$

$$
\begin{aligned}
R_X &= a_0 + \sum_{j>0} a_j\,\cos(\omega_j\,T - \alpha_j), \\
R_Y &= \sum_{j>0} b_j\,\sin(\omega_j\,T - \alpha_j), \\
R_Z &= \sum_{j>0} c_j\,\sin(\Omega_j\,T - \gamma_j).
\end{aligned}
$$

$$
\begin{aligned}
x_0 &= -\frac{a_0}{3\,(1+m^2/2)}, \\
x_j &= \frac{\omega_j\,a_j - 2\,b_j}{\omega_j\,(1-3\,m^2/2-\omega_j^2)}, \\
y_j &= \frac{(\omega_j^2+3+3\,m^2/2)\,b_j - 2\,\omega_j\,a_j}{\omega_j^2\,(1-3\,m^2/2-\omega_j^2)}, \\
z_j &= \frac{c_j}{1+3\,m^2/2-\Omega_j^2},
\end{aligned}
$$

$$
\begin{aligned}
\delta X &= x_0 + \sum_{j>0} x_j\,\cos(\omega_j\,T - \alpha_j), \\
\delta Y &= \sum_{j>0} y_j\,\sin(\omega_j\,T - \alpha_j), \\
\delta Z &= \sum_{j>0} z_j\,\sin(\Omega_j\,T - \gamma_j).
\end{aligned}
$$

Final Form, after comparing with tabulated values

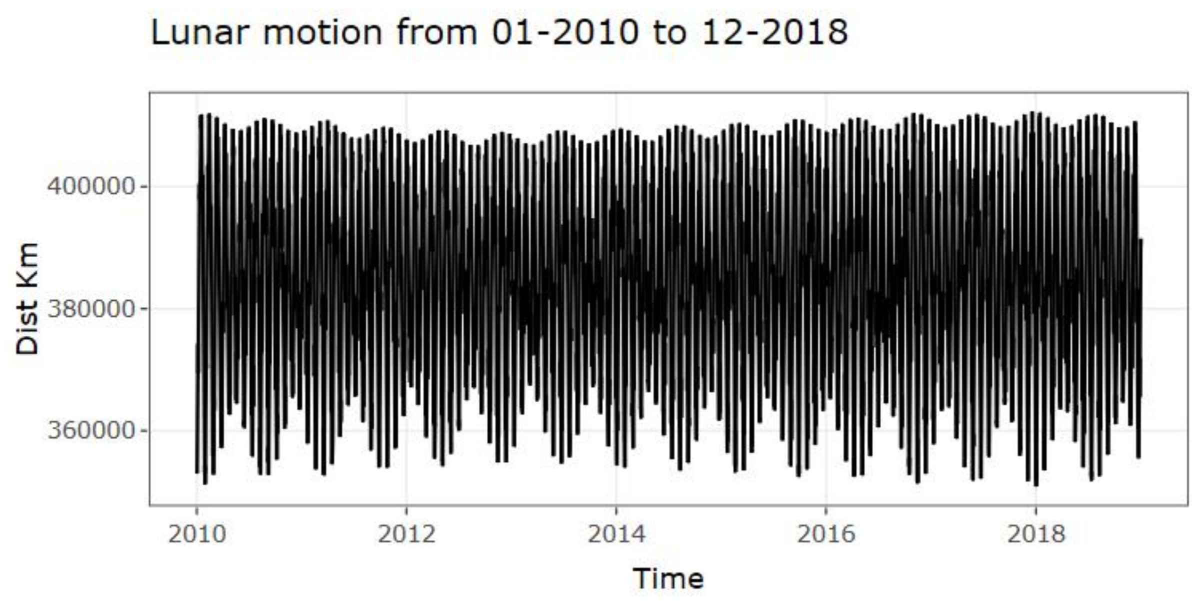
$$
\begin{aligned}
\delta X &= -\frac{1}{2}\,e^2 - \frac{1}{4}\,\iota^2 - e\,\cos[(1+c\,m^2)\,T - \alpha_0] + \frac{1}{2}\,e^2\,\cos[2\,(1+c\,m^2)\,T - 2\,\alpha_0] \\
&\quad + \frac{1}{4}\,\iota^2\,\cos[2\,(1+g\,m^2)\,T - 2\,\gamma_0] - m^2\,\cos[2\,(1-m)\,T] \\
&\quad - \frac{15}{8}\,m\,e\,\cos[(1-2\,m-c\,m^2)\,T + \alpha_0], \\
\delta Y &= 2\,e\,\sin[(1+c\,m^2)\,T - \alpha_0] + \frac{1}{4}\,e^2\,\sin[2\,(1+c\,m^2)\,T - 2\,\alpha_0] \\
&\quad - \frac{1}{4}\,\iota^2\,\sin[2\,(1+g\,m^2)\,T - 2\,\gamma_0] + \frac{11}{8}\,m^2\,\cos[2\,(1-m)\,T]
\end{aligned}
$$

$$
\begin{aligned}
\delta Z &= \iota\,\sin[(1+g\,m^2)\,T - \gamma_0] + \frac{3}{2}\,e\,\iota\,\sin[(c-g)\,m^2\,T - \alpha_0 + \gamma_0] \\
&\quad + \frac{1}{2}\,e\,\iota\,\sin[(2+c\,m^2+g\,m^2)\,T - \alpha_0 - \gamma_0] \\
&\quad + \frac{3}{8}\,m\,\iota\,\sin[(1-2\,m-g\,m^2)\,T + \gamma_0].
\end{aligned}
$$

IN DETAILS

practical solution: **pyEphem** (RK4 + trigonometric triangulation)



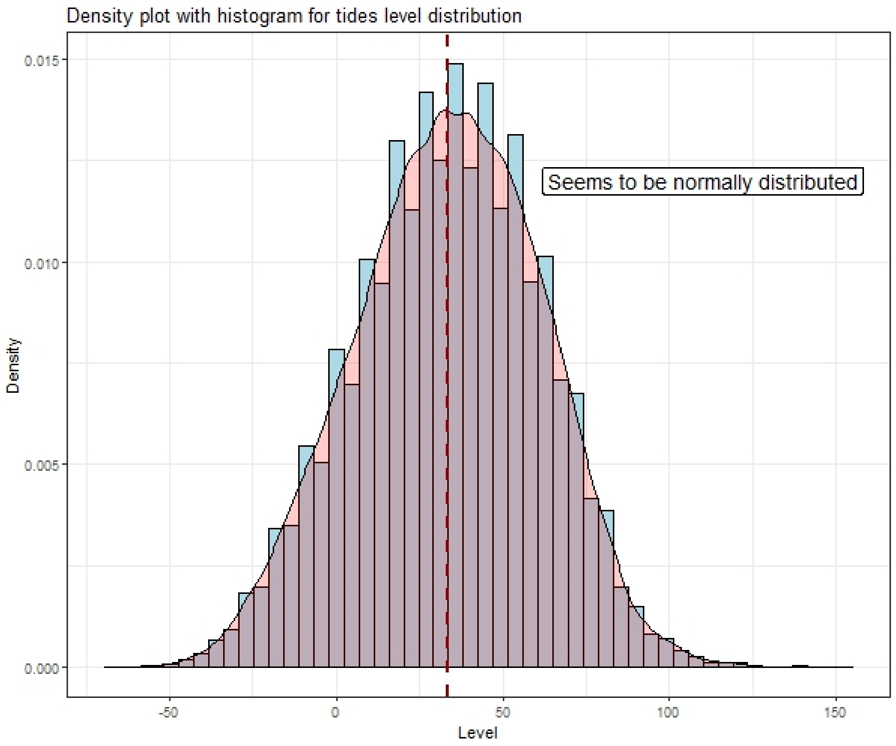
Source: https://faber6911.github.io/venice-is-

Web Viewer Terms | Privacy & Cookies

Edit

DATA INSPECTION

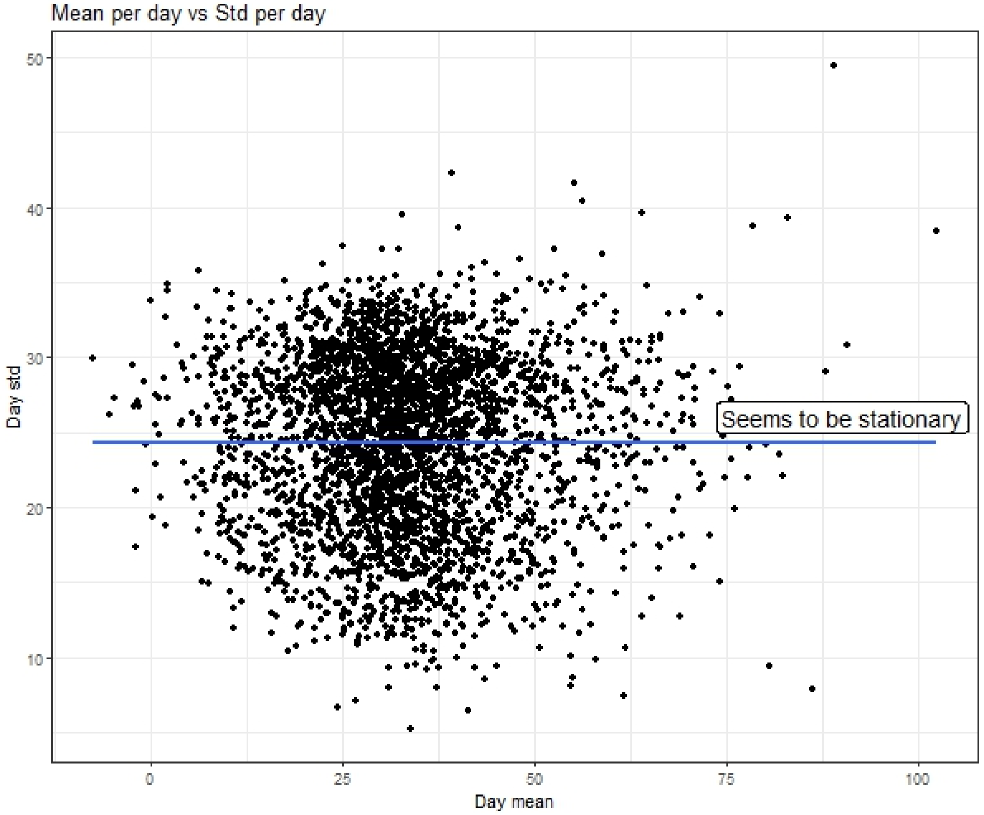
INTERESTING INSPECTION



data regarding sea level seems to be normally distributed, so from the analytic perspective the concepts of strict and weak stationarity are equivalent

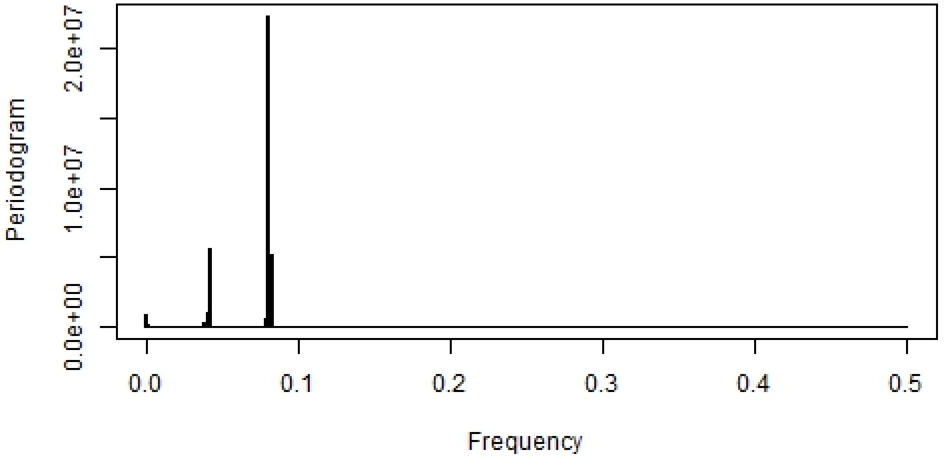
Dickey-Fuller test confirms in-mean stationarity

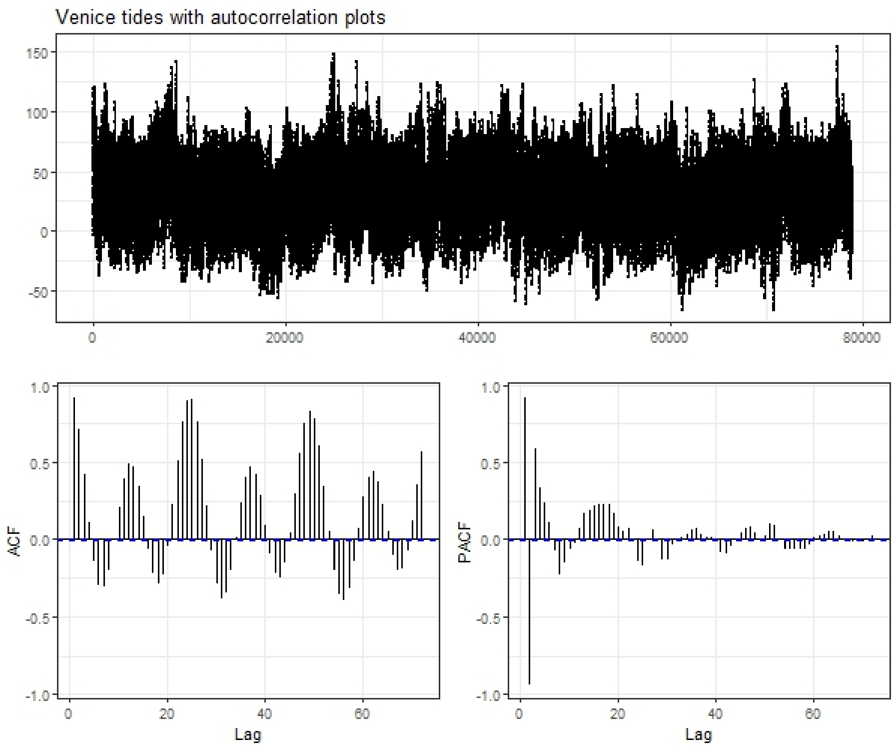
```
Value of test-statistic is: -15.6087 121.8284

Critical values for test statistics:
      1pct  5pct 10pct
tau2 -3.43 -2.86 -2.57
phi1  6.43  4.59  3.78
```



also in-variance stationarity seems to be confirmed

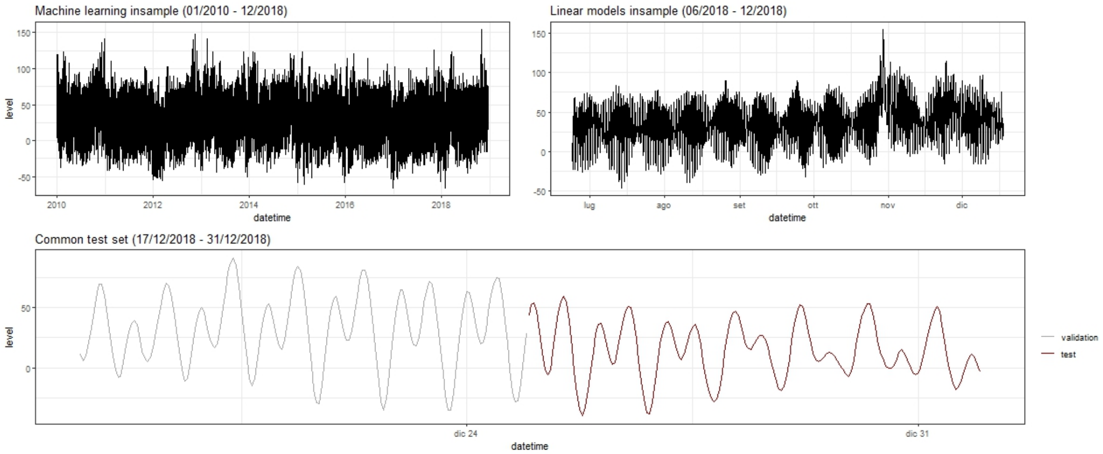
using periodogram is possible to detect frequency at 11-12 and 23-24-25 hours





for computational reasons we limited our data to the interval 2010-2018

the training set is 10 year for ML model and 6 months for linear models, the validation and test sets are 1 week each one

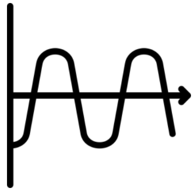


MODELS

ARIMA

two arima models using different regressors

weather data with lunar motion

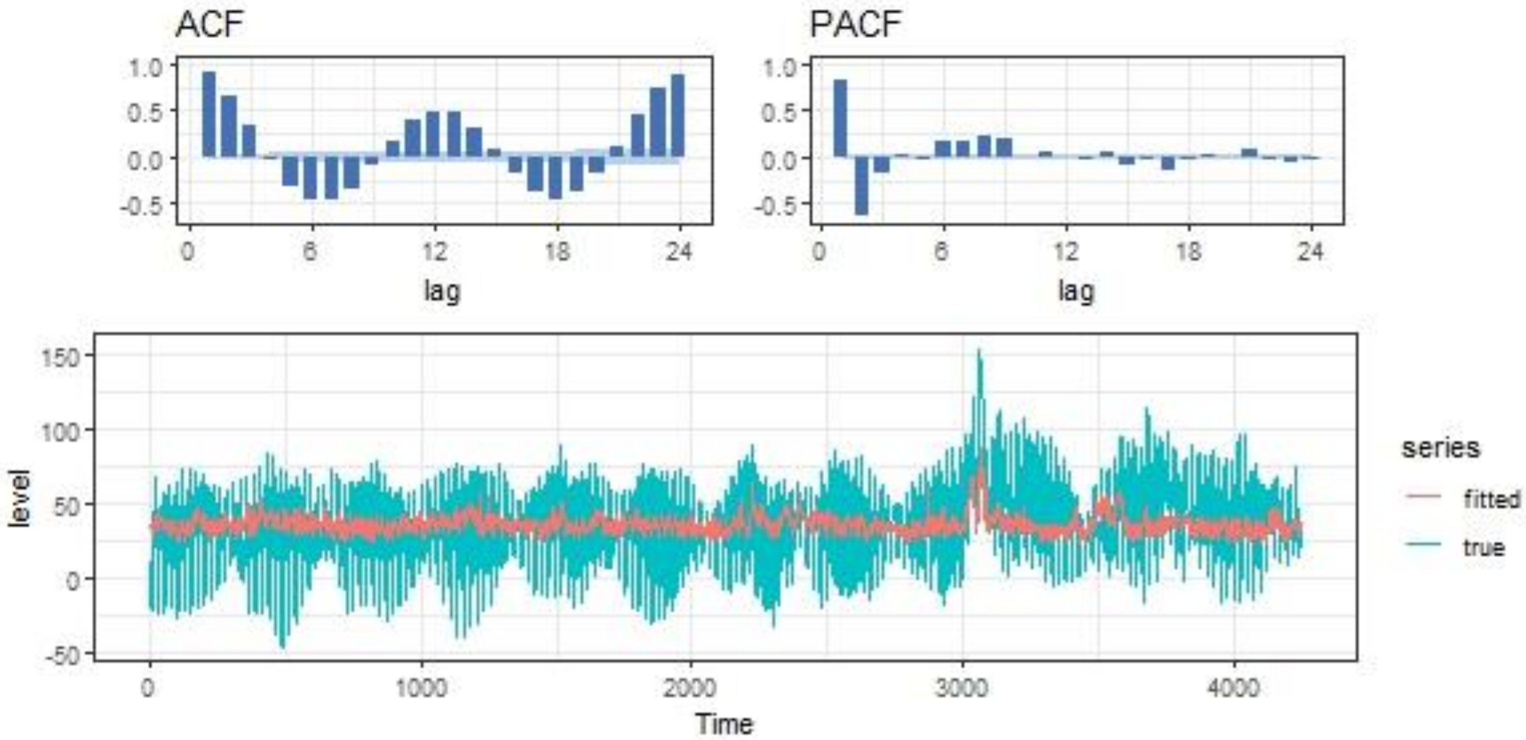


harmonics from *oce* package:

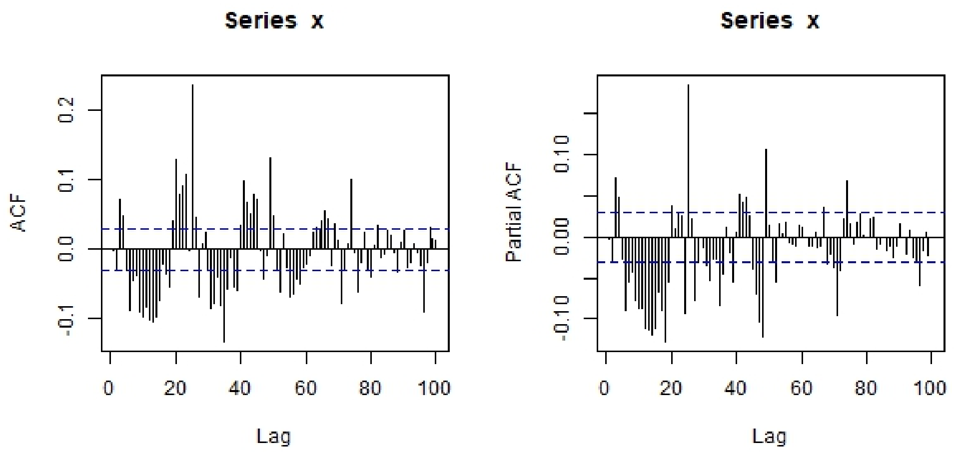
- M2, main lunar semi-diurnal with a period of ~12 hours;
- S2, main solar semi-diurnal (~12 hours);
- N2, lunar-elliptic semi-diurnal (~13 hours);
- K2, lunar-solar semi-diurnal (~12 hours);
- K1, lunar-solar diurnal (~24 hours);
- O1, main lunar diurnal (~26 hours);
- SA, solar annual (~24*365 hours);
- P1, main solar diurnal (24 hours)

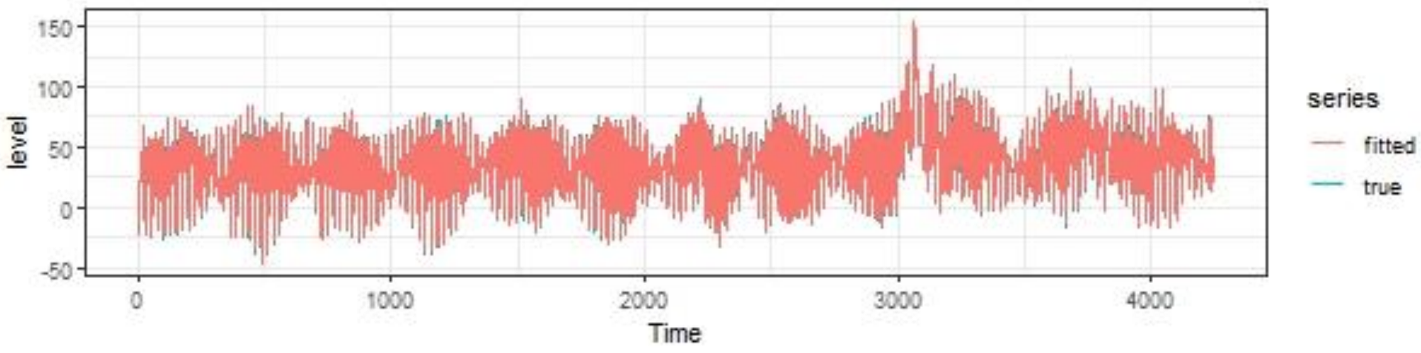
first model exploit weather data in combination with the lunar motion

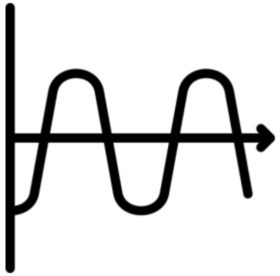
initial fitting performances



After several attemps, following the represented Lag on ACF and PACF plots in combination with the value of the AICc and the Mean Absolute Percentage Error (MAPE), a highly parameterized model has been reached with the form (3,1,3)(1,1,3)[24]







harmonics in detail

Insert Web Page

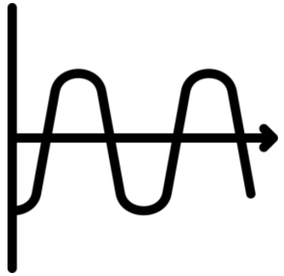
This app allows you to insert secure web pages starting with https:// into the slide deck. Non-secure web pages are not supported for security reasons.

Please enter the URL below.

https:// faber6911.github.io/venice-is-drowning/report/imgs/components.html

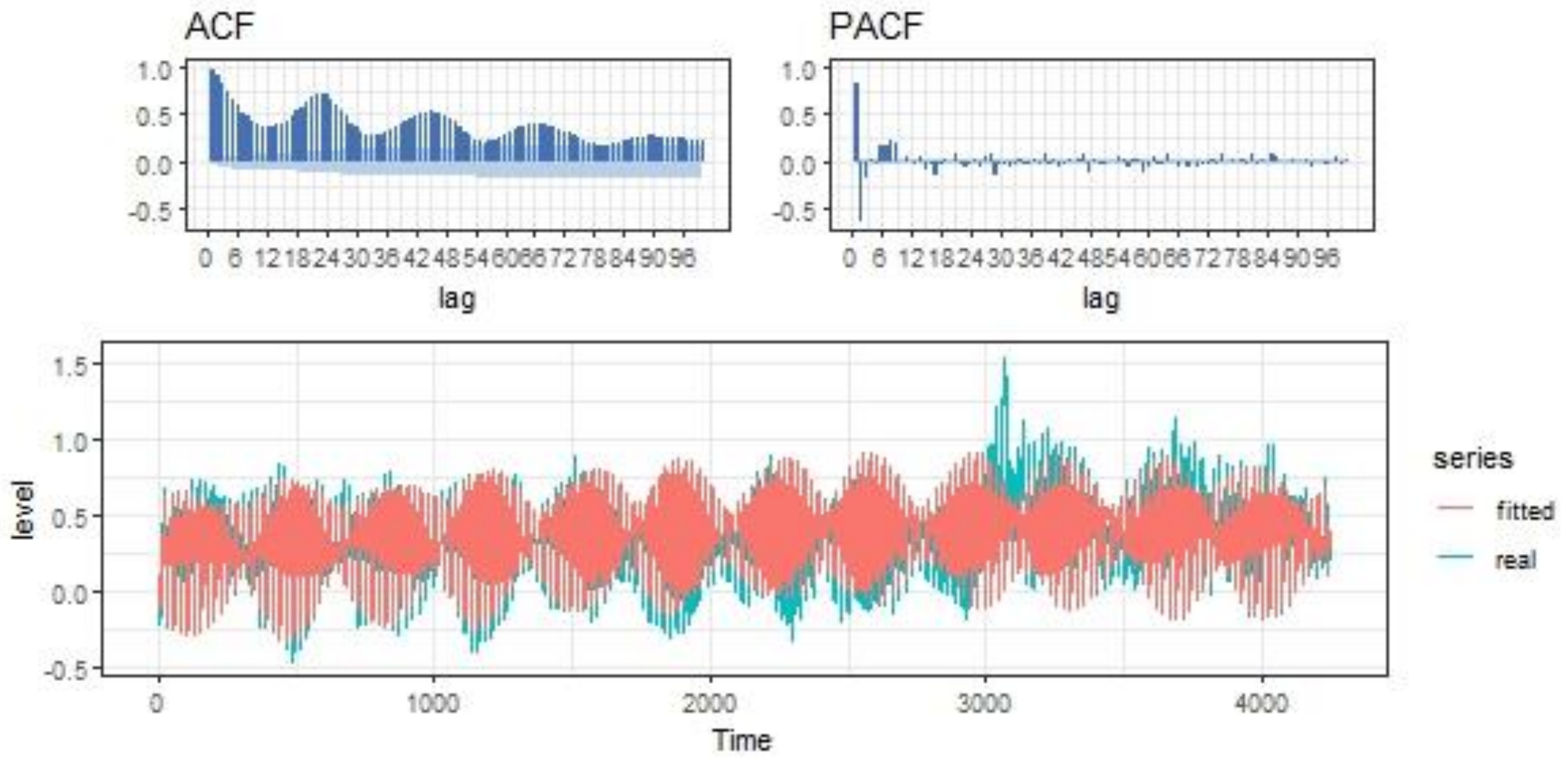
Note: Many popular websites allow secure access. Please click on the preview button to ensure the web page is accessible.

Web Viewer Terms | Privacy & Cookies Preview

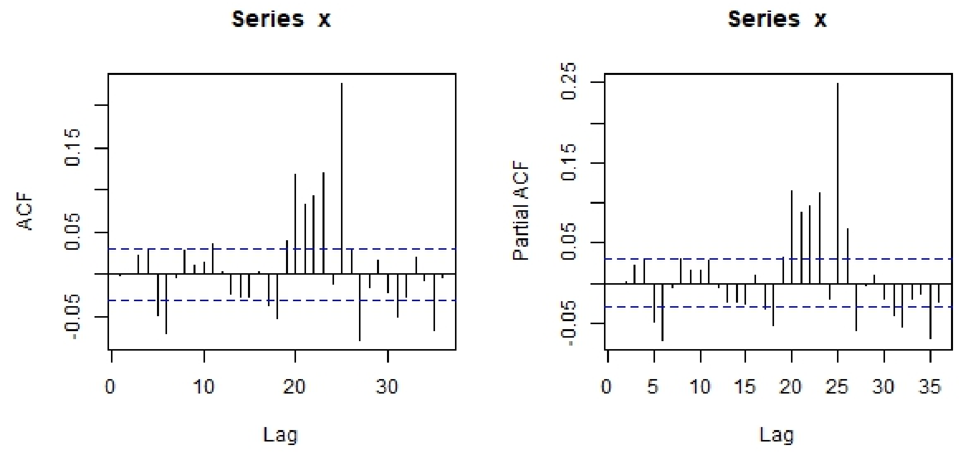


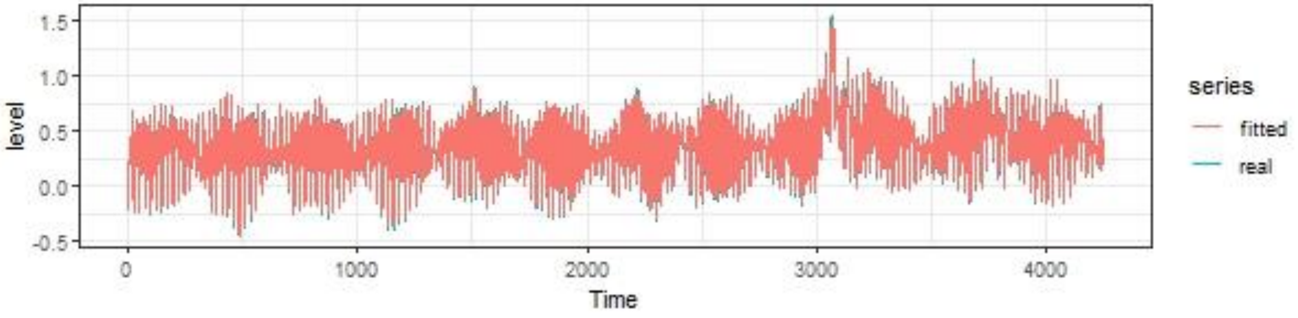
second model exploit harmonics as regressors

initial fitting perfomances are significantly improved



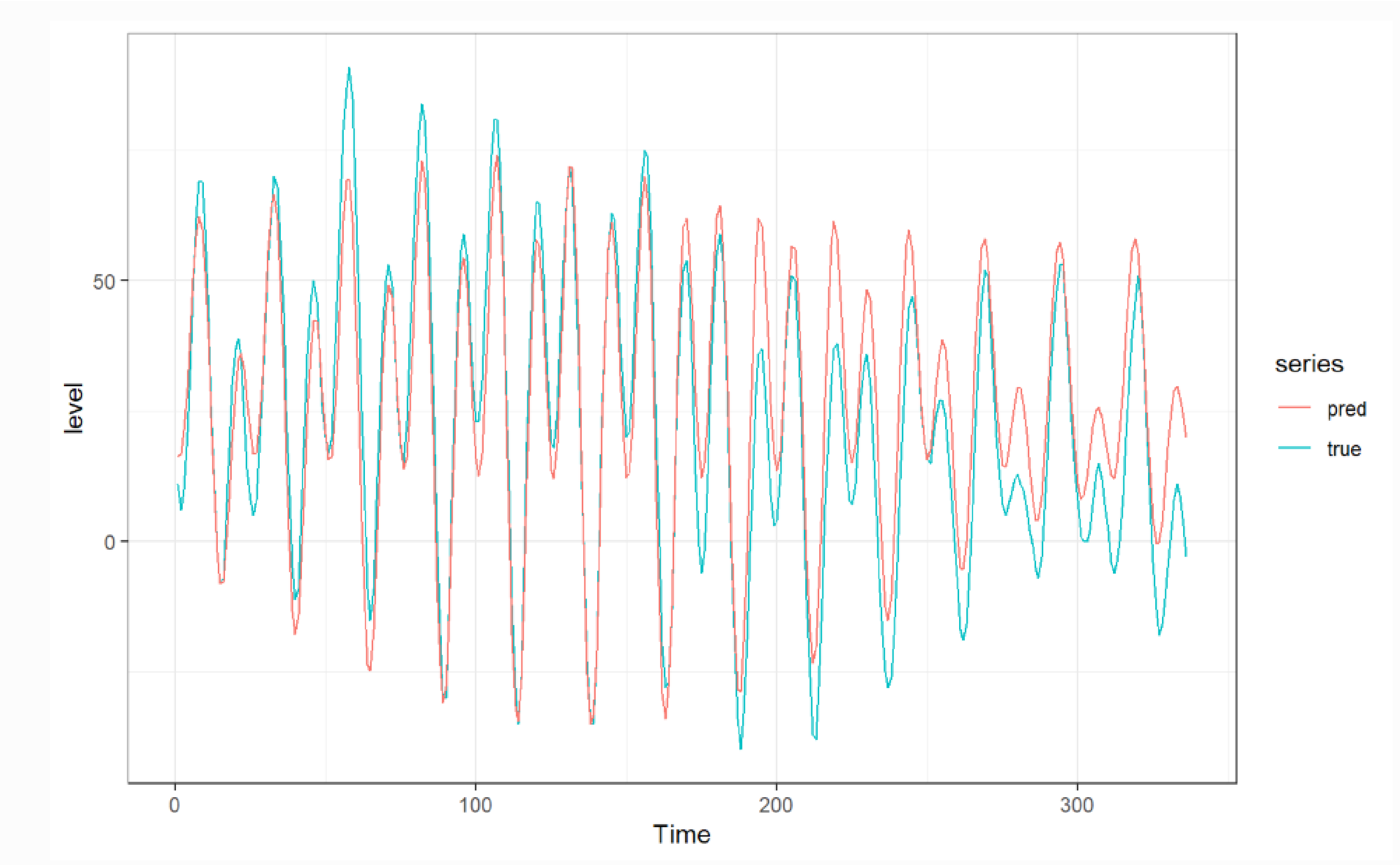
Also in this case, following the observations of the autocorrelation plots and the trend of the AICc we proceeded to insert the autoregressive and moving average components until obtaining the final model (3,0,2)(1,0,0)[24] with drift





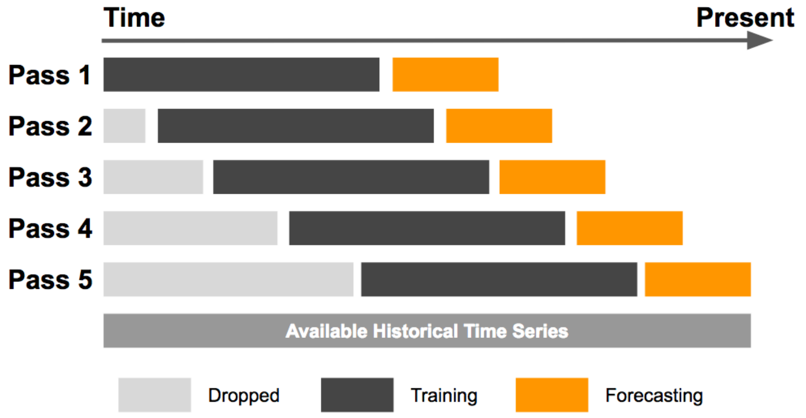
UCM

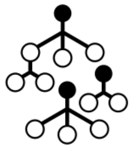
The best performing model was found (after several attemps) to be the one composed of harmonics inserted directly as components together with a trend component, specifically a random walk



MACHINE LEARNING

supervised problem





initial comparison

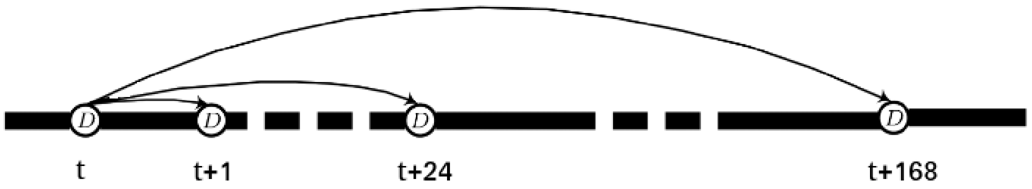
Table 1: Initial explorative comparison of the performances of the machine learning models for the 24-step ahead predictions iterated over real data and averaged over the validation set.

	Random Forest	GRU	LSTM
RMSE (cm)	12.72	8.34	6.96

LSTM

the model

```python
def make_model(n_input, n_features, verbose = False, multi = True, use_CuDNNLSTM = True,
               loss = "mse", metrics = ["mae", "mape"], lr = 0.001):
    K.clear_session()
    LSTM_layer = LSTM if not use_CuDNNLSTM else CuDNNLSTM
    opt = Adam (lr = lr )

    model = Sequential()
    model.add(LSTM_layer(512, input_shape=(n_input, n_features), return_sequences=True))
    model.add(BatchNormalization())
    model.add(LeakyReLU())
    model.add(Dropout(rate = 0.4))

    for i in range(1):
        model.add(LSTM_layer(256, return_sequences = False))
        model.add(BatchNormalization())
        model.add(LeakyReLU())
        model.add(Dropout(rate = 0.3))

    model.add(Dense(128))
    model.add(LeakyReLU())
    model.add(Dropout(rate = 0.2))

    model.add(Dense(1))
    return model
```

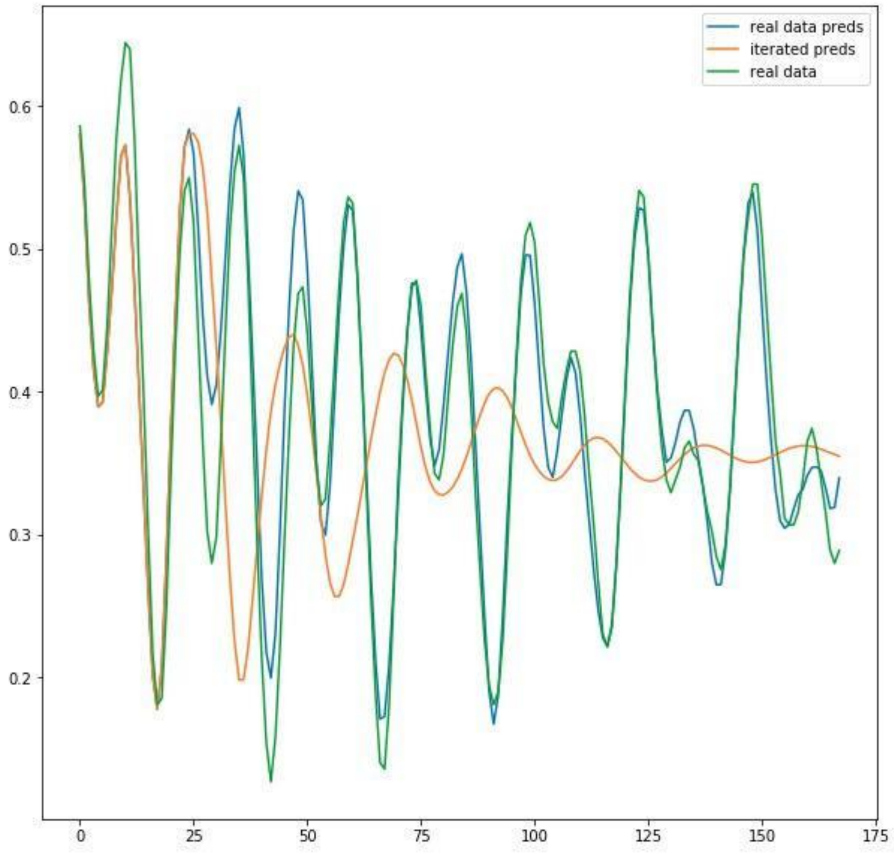


Window size depends on the prediction type:

1h → 2 h
24h → 1 day
168h → 7 day

Different evaluations (RMSE, MAPE):

- One shot
- Averaged (real data)
- Averaged (predictions)



RESULTS

RESULTS – interactive plot

Insert Web Page

This app allows you to insert secure web pages starting with https:// into the slide deck. Non-secure web pages are not supported for security reasons.

Please enter the URL below.

https:// faber6911.github.io/venice-is-drowning/report/imgs/results.html

Note: Many popular websites allow secure access. Please click on the preview button to ensure the web page is accessible.

Web Viewer Terms | Privacy & Cookies Preview

RESULTS

metrics : **MAPE** (also RMSE, MAE)

linear models generally **better** performing (“gaussian” characteristics of the process?)

mod2_ar benefits from the **harmonics** regressor

lstm has **precise** one-shot predictions (still benefits from external regressors)

MAPE (%)	mod1_ar		mod2_ar		ucm1		lstm1			lstm2		
	it.	punct.	it.	punct.	it.	punct.	it.(real)	it.(pred)	punct.	it.(real)	it.(pred)	punct.
1-step	0.11	0.93	**0.09**	0.41	0.12	0.46	1.76	18.7	0.49	2.22	13.54	**0.16**
24-steps	0.66	1.7	**0.51**	0.71	0.59	0.34	6.11	20.72	0.95	5.39	19.22	**0.22**
168-steps	2.89	0.71	2.45	**0.36**	**2.16**	0.42	15.68	-	10.19	9.12	-	9.03

CONCLUSIONS AND FUTURE WORKS

CONCLUSIONS

tested different linear (ARIMA, UCM) and non-linear (LSTM) models

1, 24, 168 steps ahead predictions

interesting results with linear models for average predictions over a week

non-linear models more precise for one-shot predictions

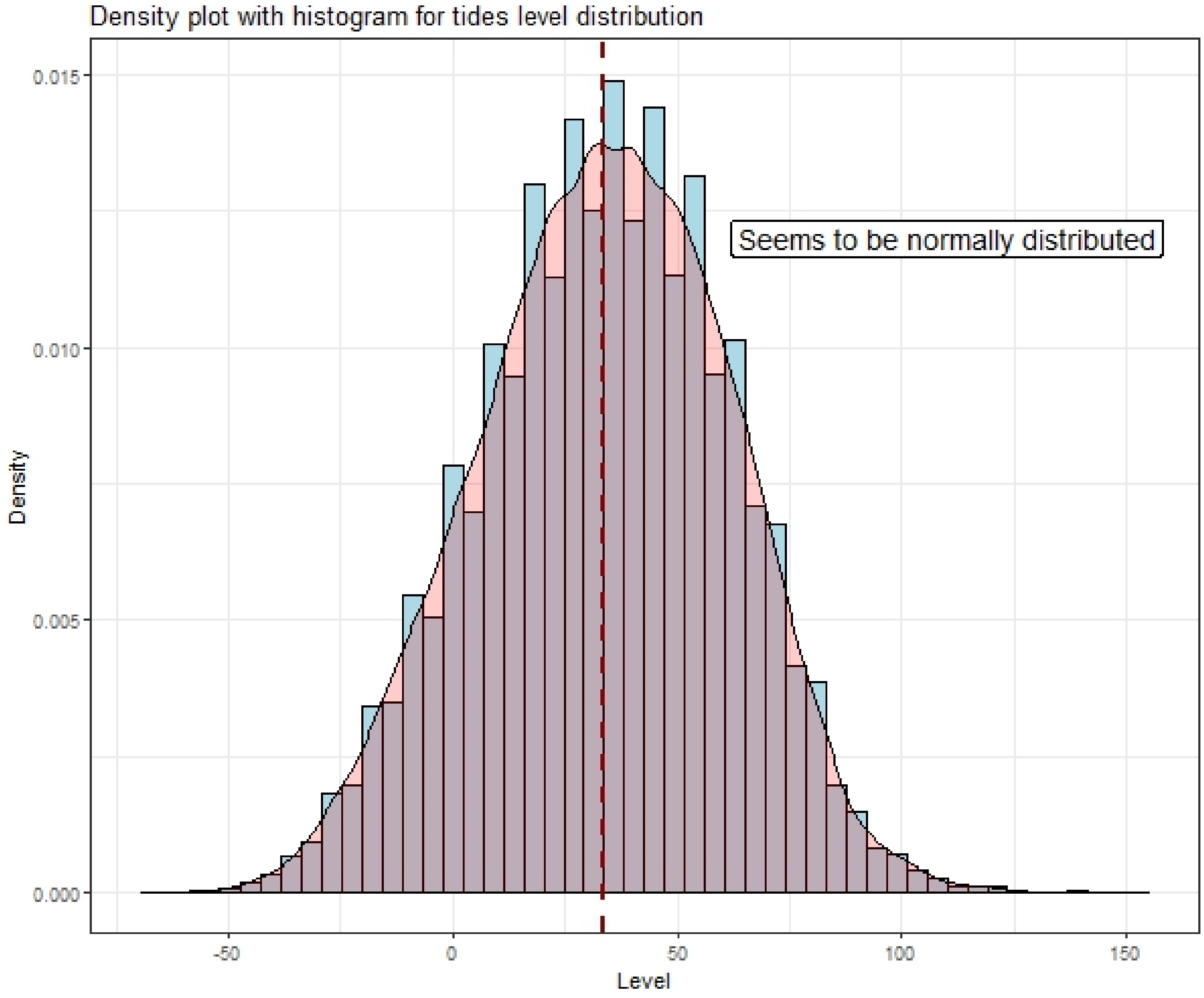
IMPROVEMENTS

more computational power

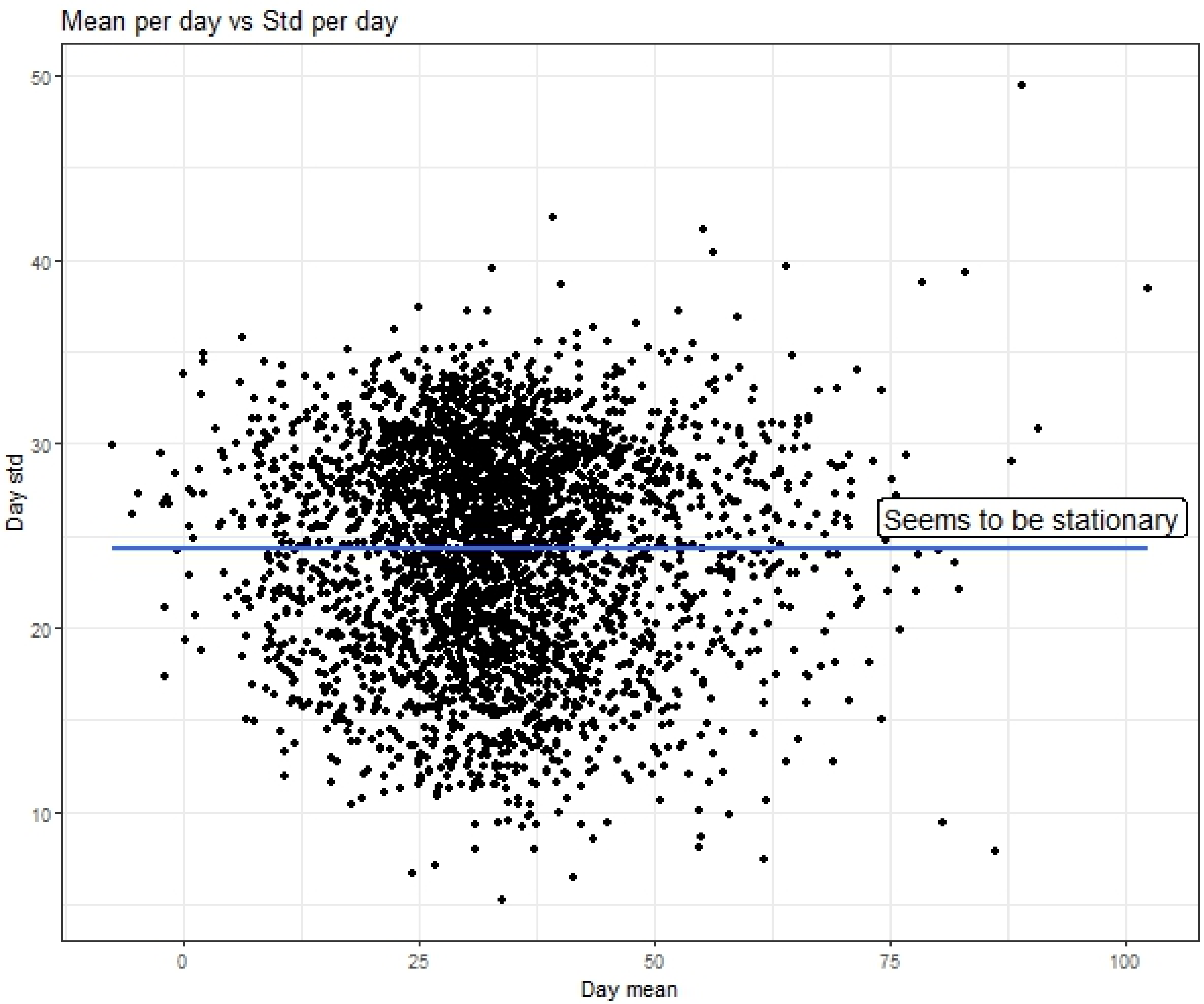
better physics modeling (i.e. hydrodynamics)

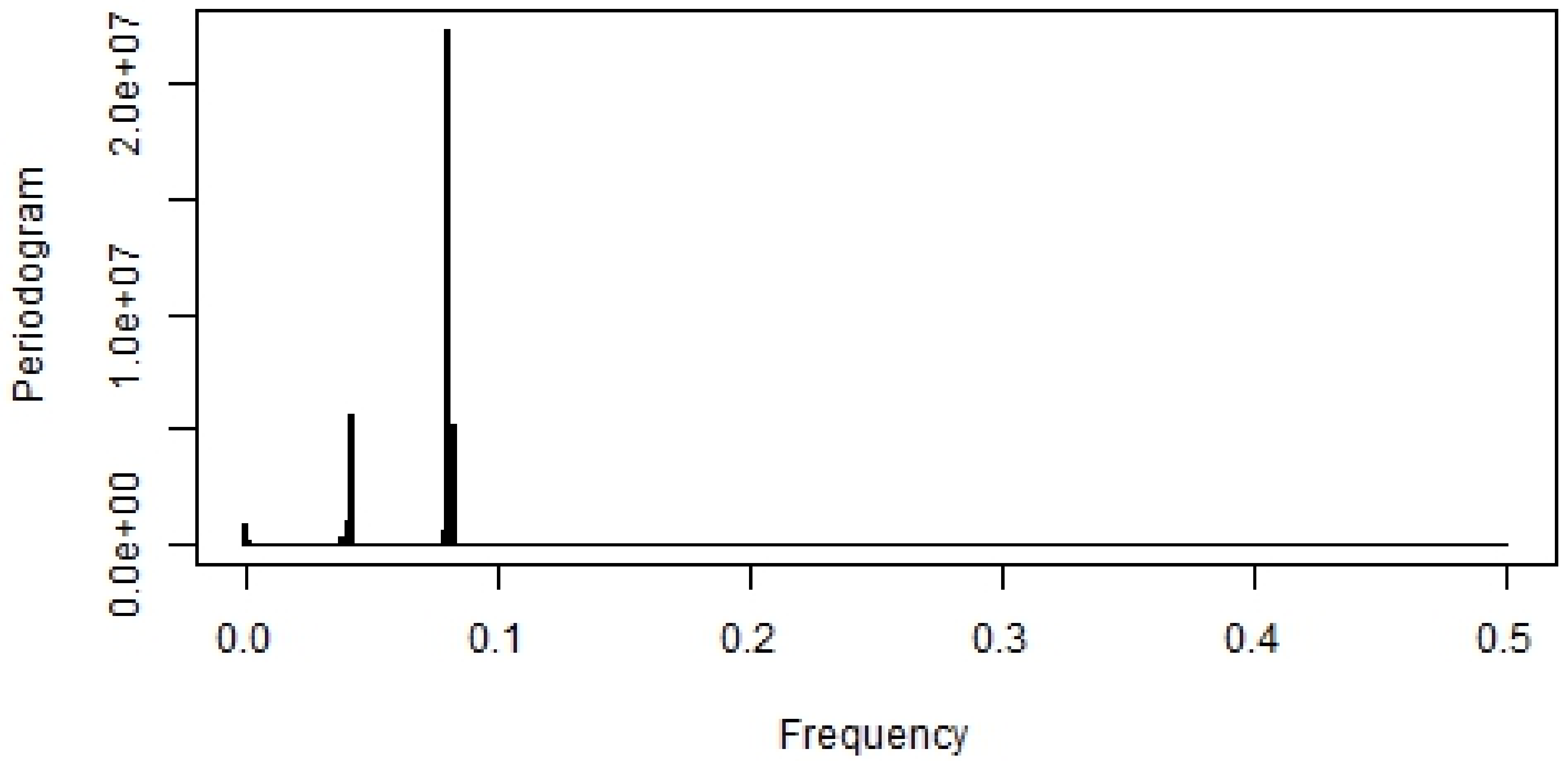
more accurate weather data

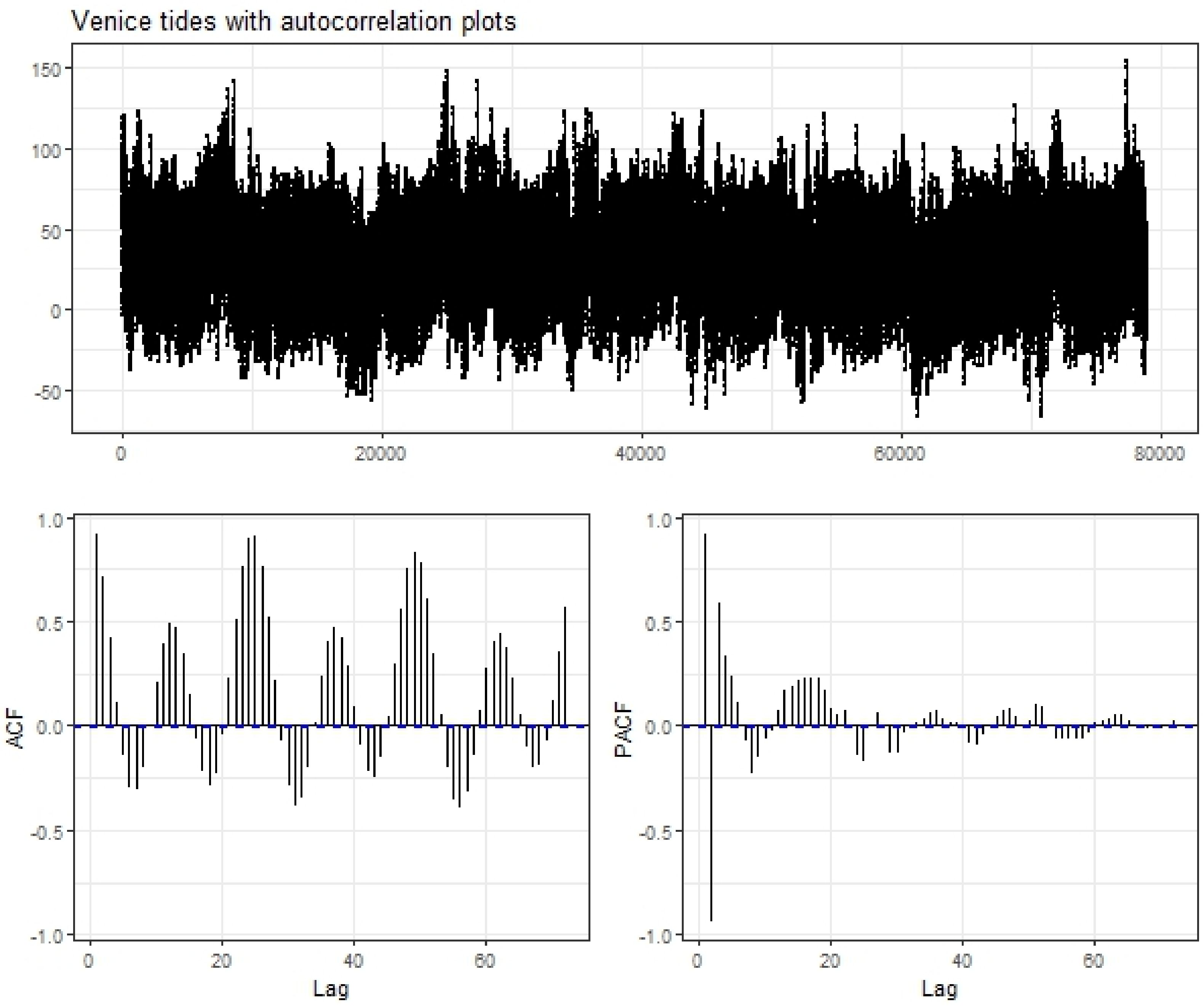
THANK YOU


Density plot with histogram for tides level distribution
Seems to be normally distributed
0.015
0.010
0.005
0.000
Density
-50
0
50
100
150
Level

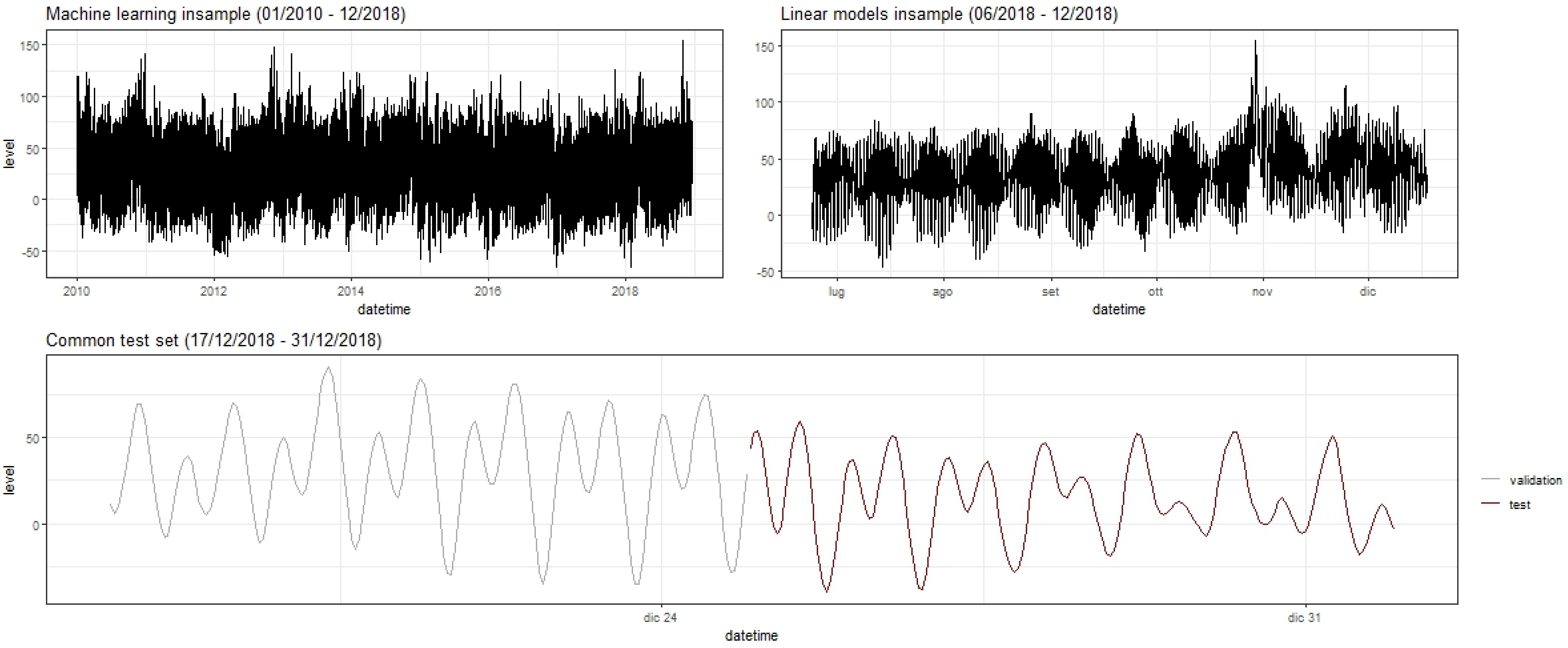
```
Value of test-statistic is: -15.6087 121.8284

Critical values for test statistics:
      1pct  5pct 10pct
tau2 -3.43 -2.86 -2.57
phi1  6.43  4.59  3.78
```


Mean per day vs Std per day
Seems to be stationary
Day std
Day mean
50
40
30
20
10
0
25
50
75
100


Periodogram
0.0e+00
1.0e+07
2.0e+07
0.0
0.1
0.2
0.3
0.4
0.5
Frequency



Venice tides with autocorrelation plots
150
100
50
0
-50
0
20000
40000
60000
80000
ACF
1.0
0.5
0.0
-0.5
-1.0
0
20
40
60
Lag
PACF
Lag


Machine learning insample (01/2010 - 12/2018)
level
150
100
50
0
-50
2010
2012
2014
2016
2018
datetime
Linear models insample (06/2018 - 12/2018)
level
150
100
50
0
-50
lug
ago
set
ott
nov
dic
datetime
Common test set (17/12/2018 - 31/12/2018)
level
50
0
dic 24
dic 31
datetime
validation
test