Acea Smart Water Analytics

Можете ли вы помочь сохранить "голубое золото", используя данные для прогнозирования наличия воды?

Выполнили студенты:

Волянский Юлиан, гр. Б17-565

Худоярова Анастасия, гр. Б17-505

Шарафиев Родион, гр. Б17-503

Описание

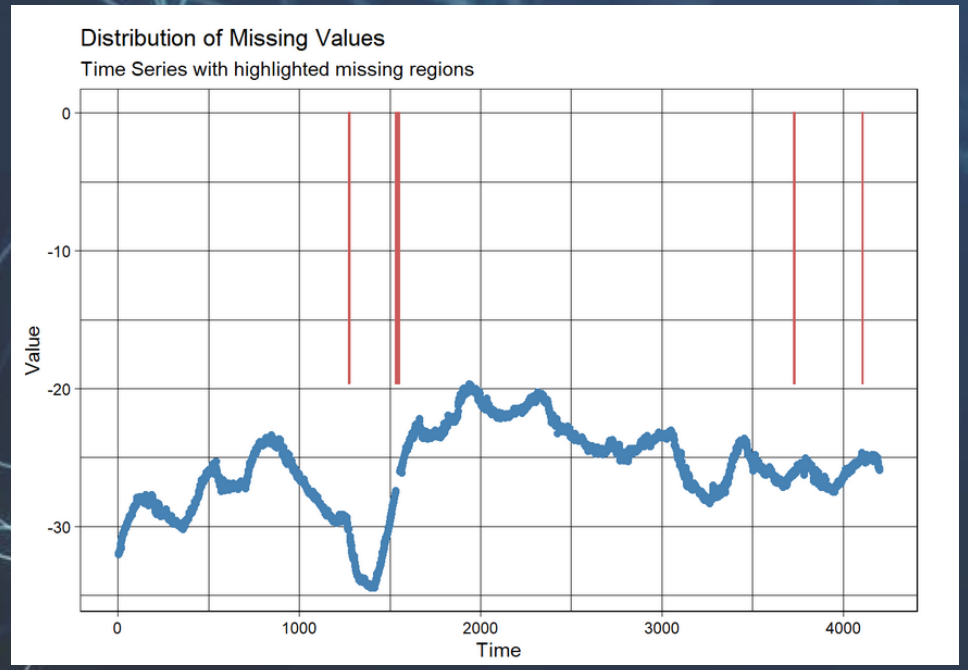
Задача: для каждого водоёма понять, что влияет на его водообеспеченность.
Временной интервал прогноза: зависит от датасета
Результат: построение математических моделей, которые прогнозируют требуемые параметры.

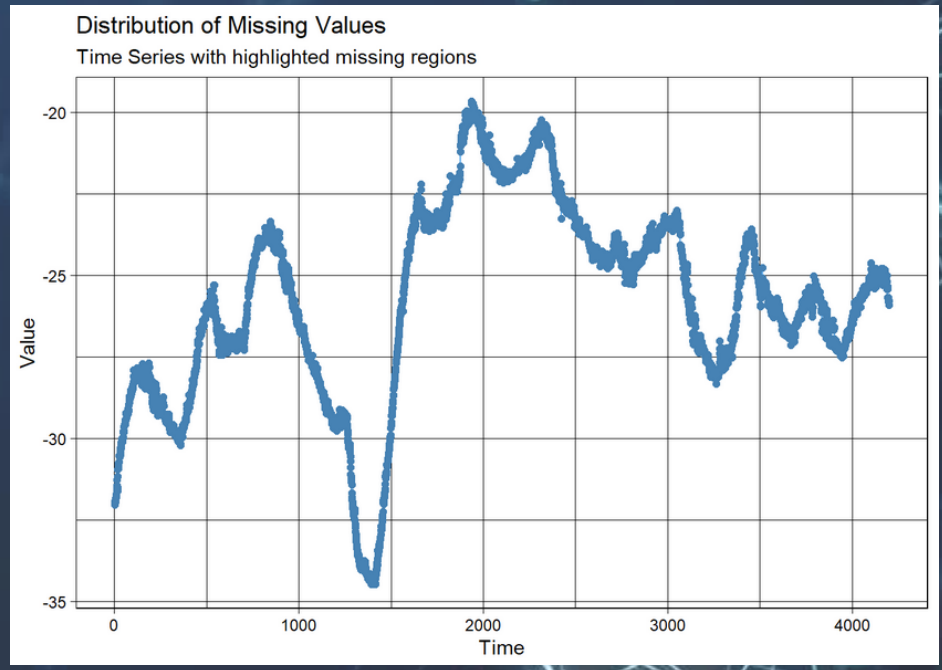
Этапы работы

- Подготовка данных
- Получения трендов, сезонности всех характеристик
- Общий анализ и выявление общих закономерностей и особенностей
- Проверка стационарности данных
- Приведение данных в случае необходимости к нужному виду.
- Построение модели
- Тестирование
- Прогнозирование

Petrigano

Подготовка данных:
удаление NaN и дополнение пропусков с помощью интерполяции значений.

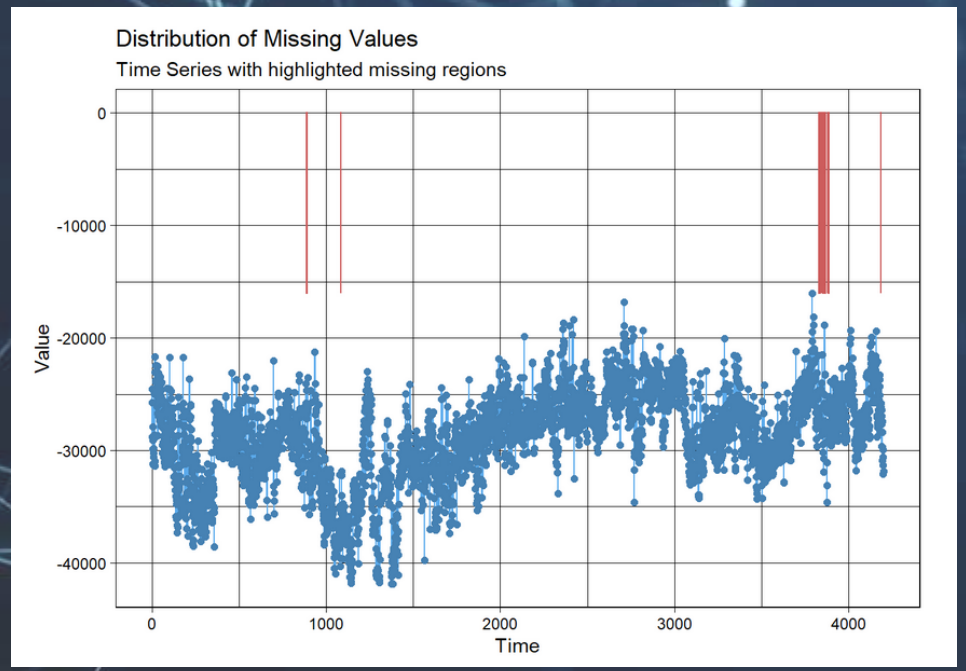


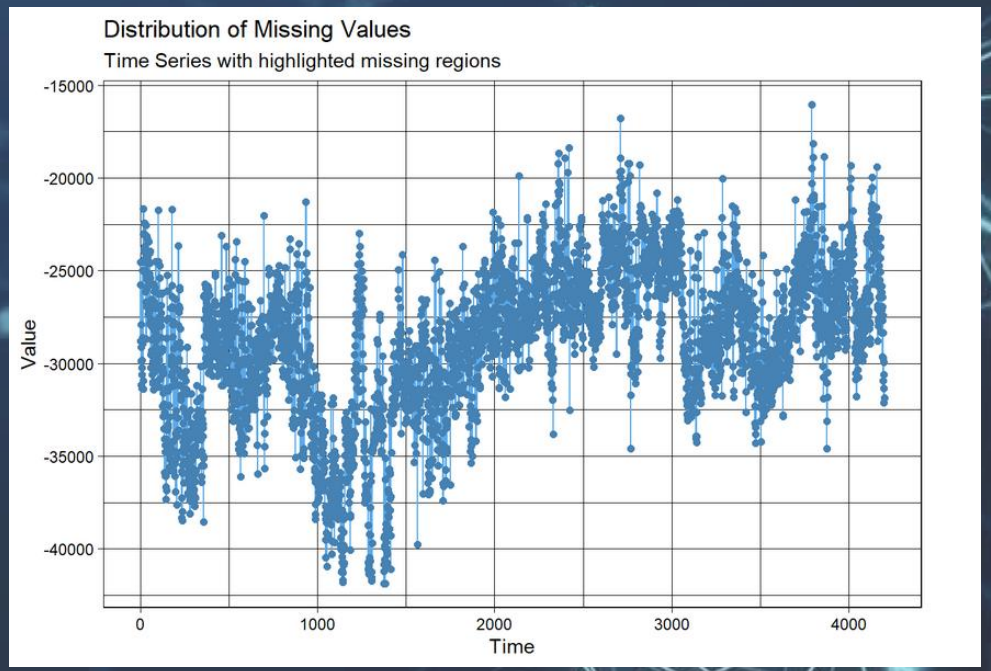


Depth

Petrigano

Подготовка данных:
удаление NaN и дополнение пропусков с помощью интерполяции значений.

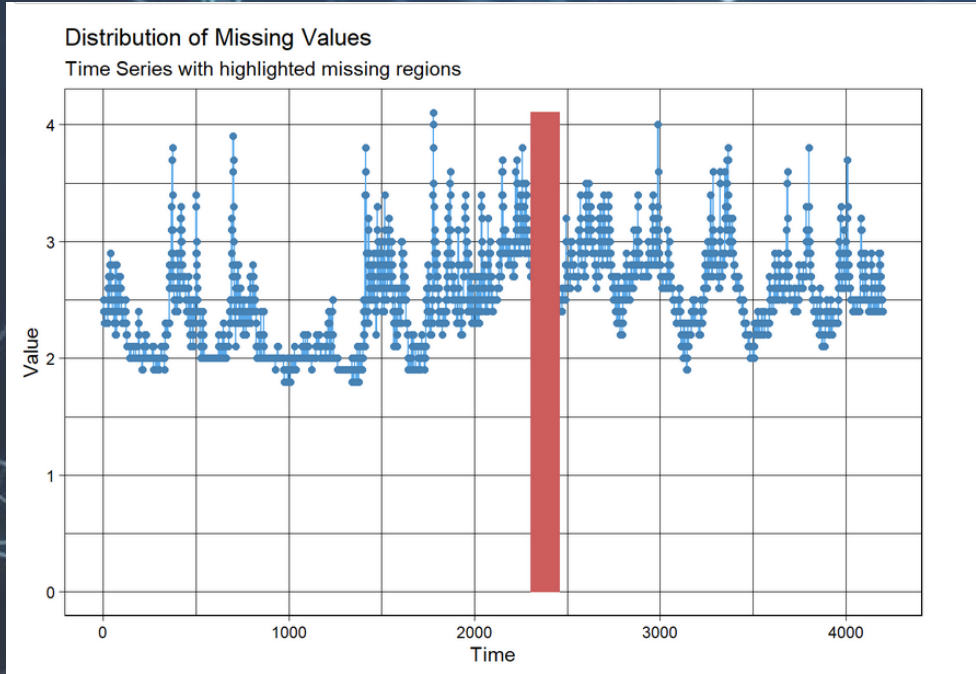


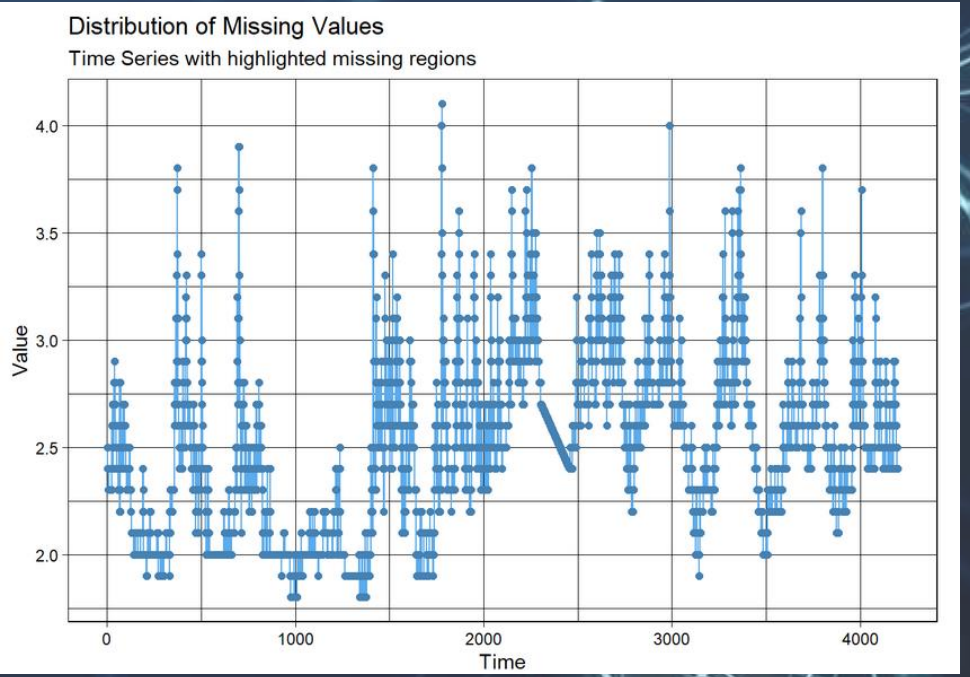


Volume

Petrigano

Подготовка данных:
удаление NaN и дополнение пропусков с помощью интерполяции значений.





Hydrometry

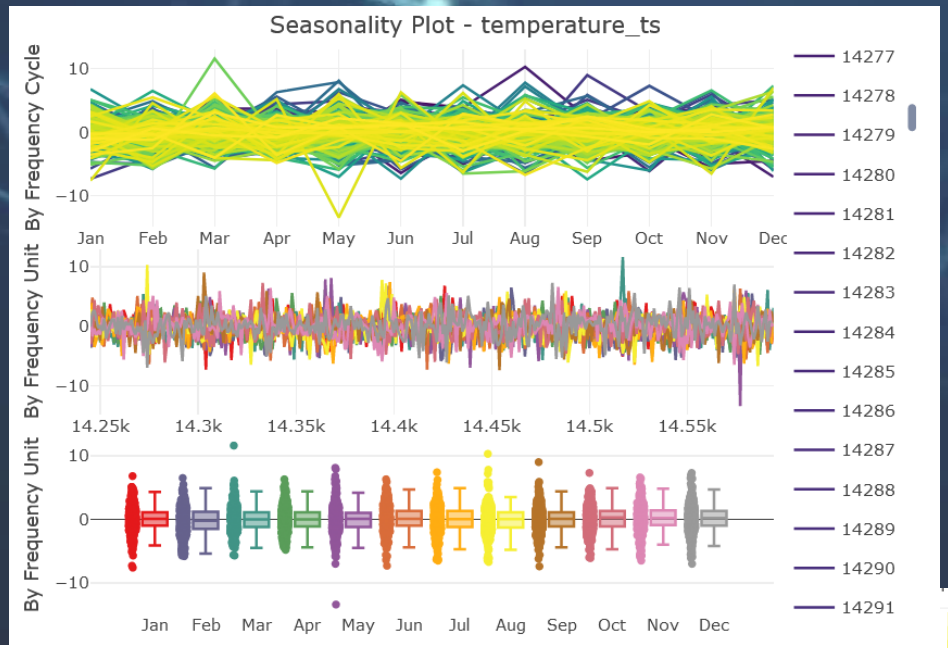
Petrigano

Подготовка данных:
Восстановление хронологического порядка

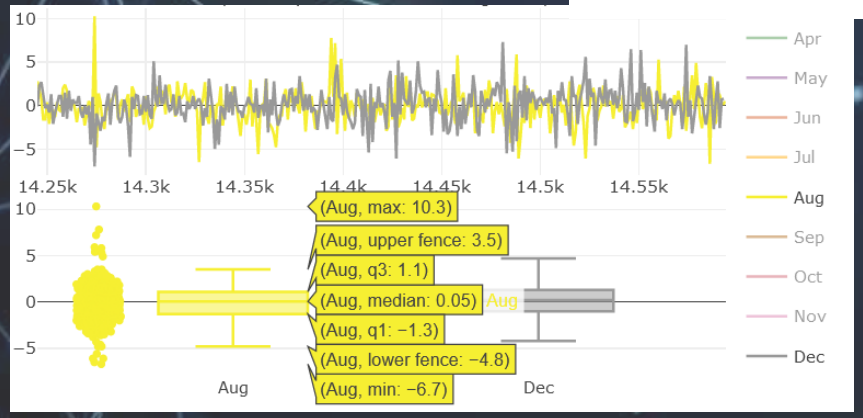
```
df <- df[order(df$Date), ]
df['Time_Interval'] = df$Date - shift(df$Date, n=1, fill=NA, type="lag")

days <- df$Time_Interval


for(i in days)
{

  if(isTRUE(i > 1) || is.null(i))
  {
    print(i)
  }

}
```

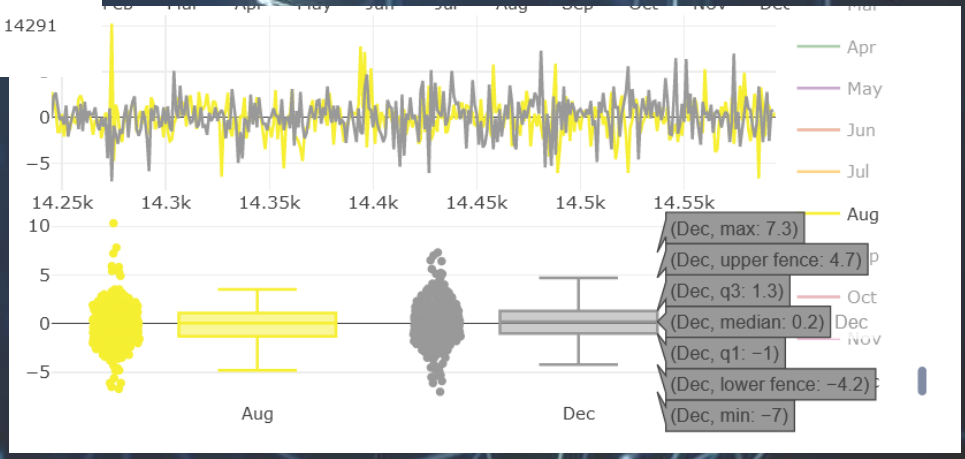
Petrigano

Декомпозиция временных рядов. Общие выводы



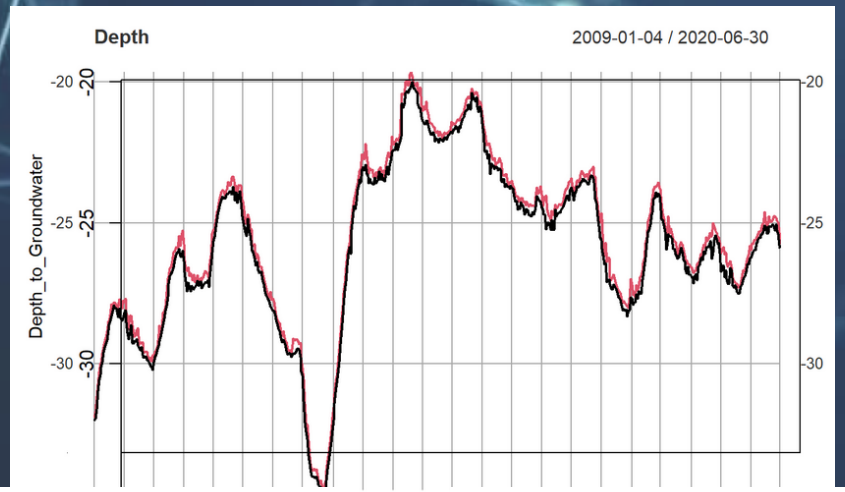
Так для всех параметров

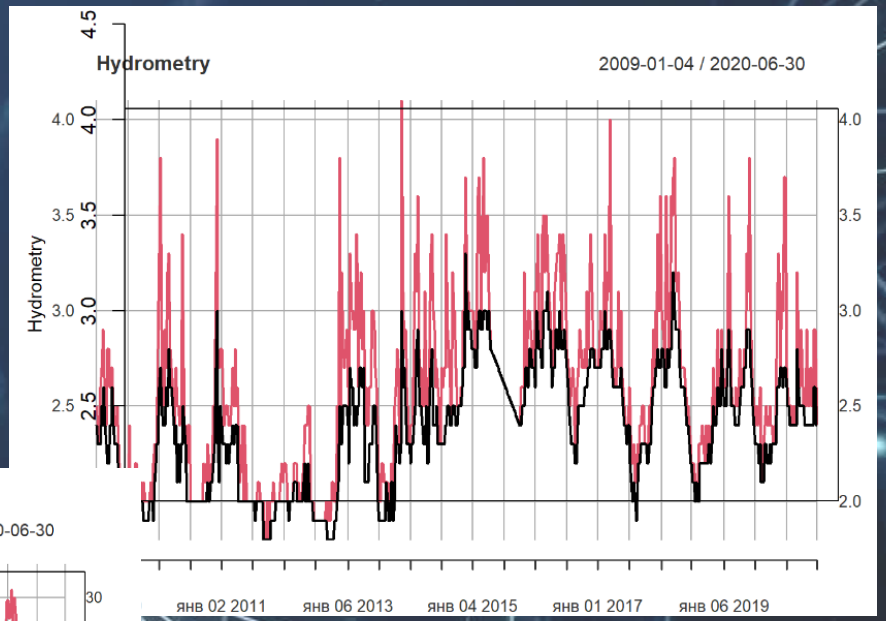


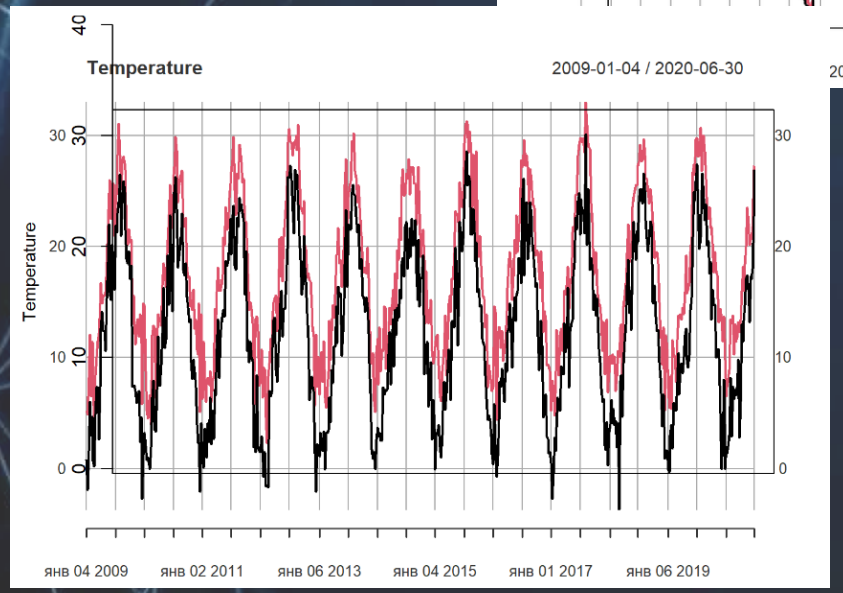


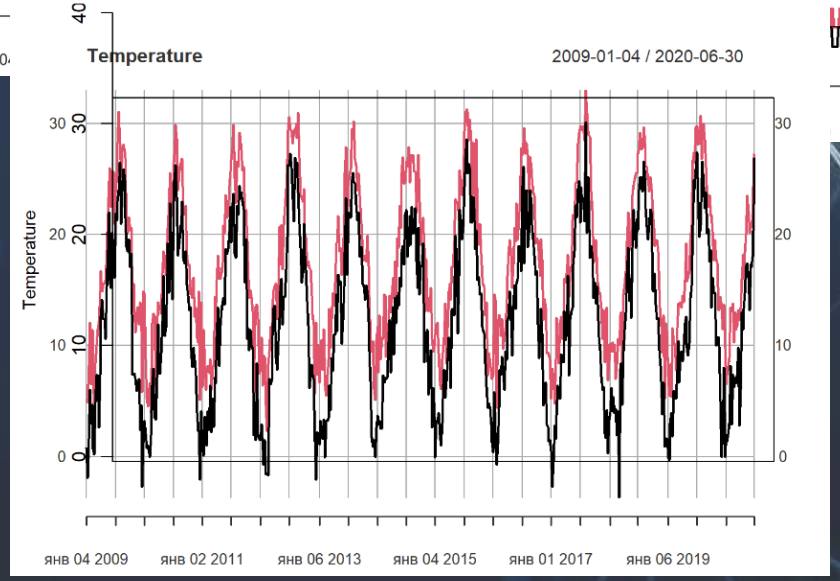
Petrigano

Ресемлирование









Petrigano

Проверка на стационарность

До преобразований

```
adfuller(train_without_seasonal)
```

```
(-1.7530523252277812,
 0.4040061812438108,
 30,
 4147,
 {'1%': -3.431927852028984,
  '5%': -2.862237208417471,
  '10%': -2.5671411303007297},
 -6466.6340148344425)
```

```
kpss(train_without_seasonal)
```

```
(4.8498423005402,
 0.01,
 31,
 {'10%': 0.347, '5%': 0.463, '2.5%': 0.574, '1%': 0.739})
```

1. **Убрали годовую сезонность**
2. **Убрали недельную сезонность**
3. **Подняли ряд до нуля с помощью среднего значения**
4. **Дифференцировали ряд**

До преобразований

```
adfuller(train_diff)
```

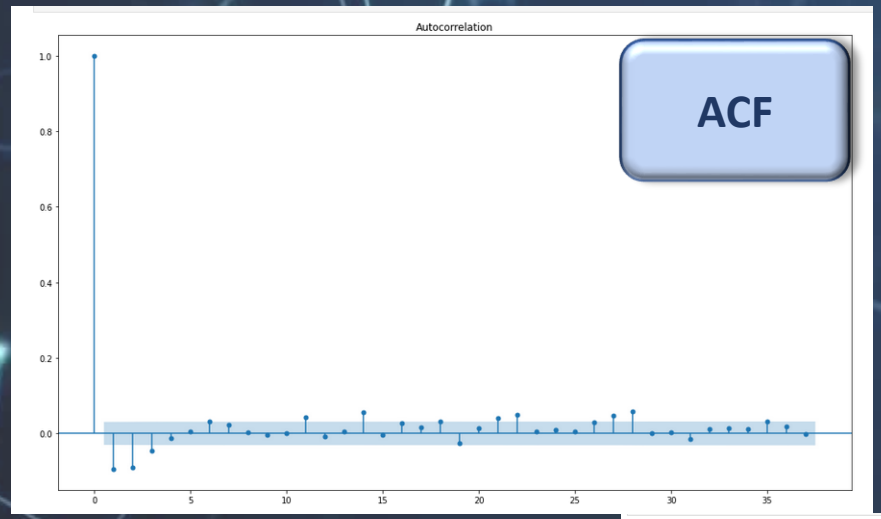
```
(-7.049388885102958,
 5.577384465690022e-10,
 29,
 4147,
 {'1%': -3.431927852028984,
  '5%': -2.862237208417471,
  '10%': -2.5671411303007297},
 -6463.0131426843)
```

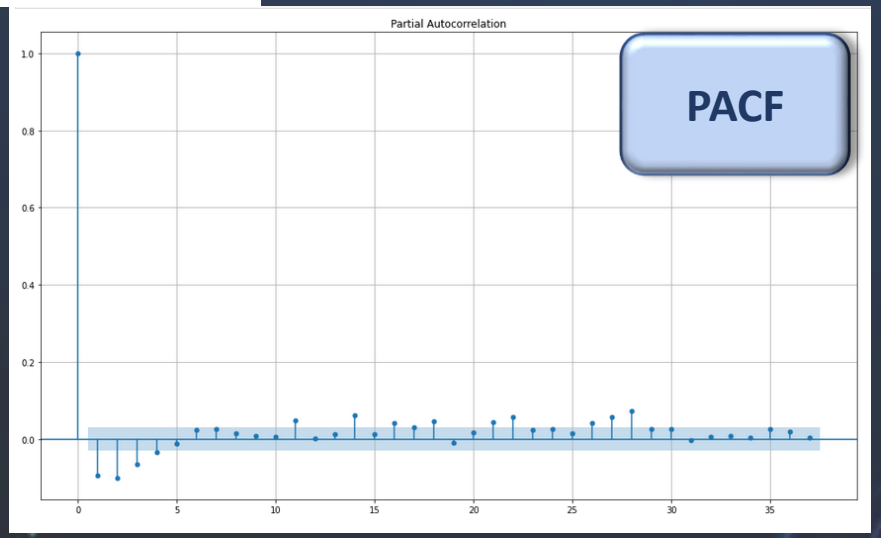
```
kpss(train_diff)
```

```
(0.41039939618941784,
 0.07267267405628541,
 31,
 {'10%': 0.347, '5%': 0.463, '2.5%': 0.574, '1%': 0.739})
```

Petrigano

Построение модели





Выбор лучшей по AIC, BIC

SARIMAX Results

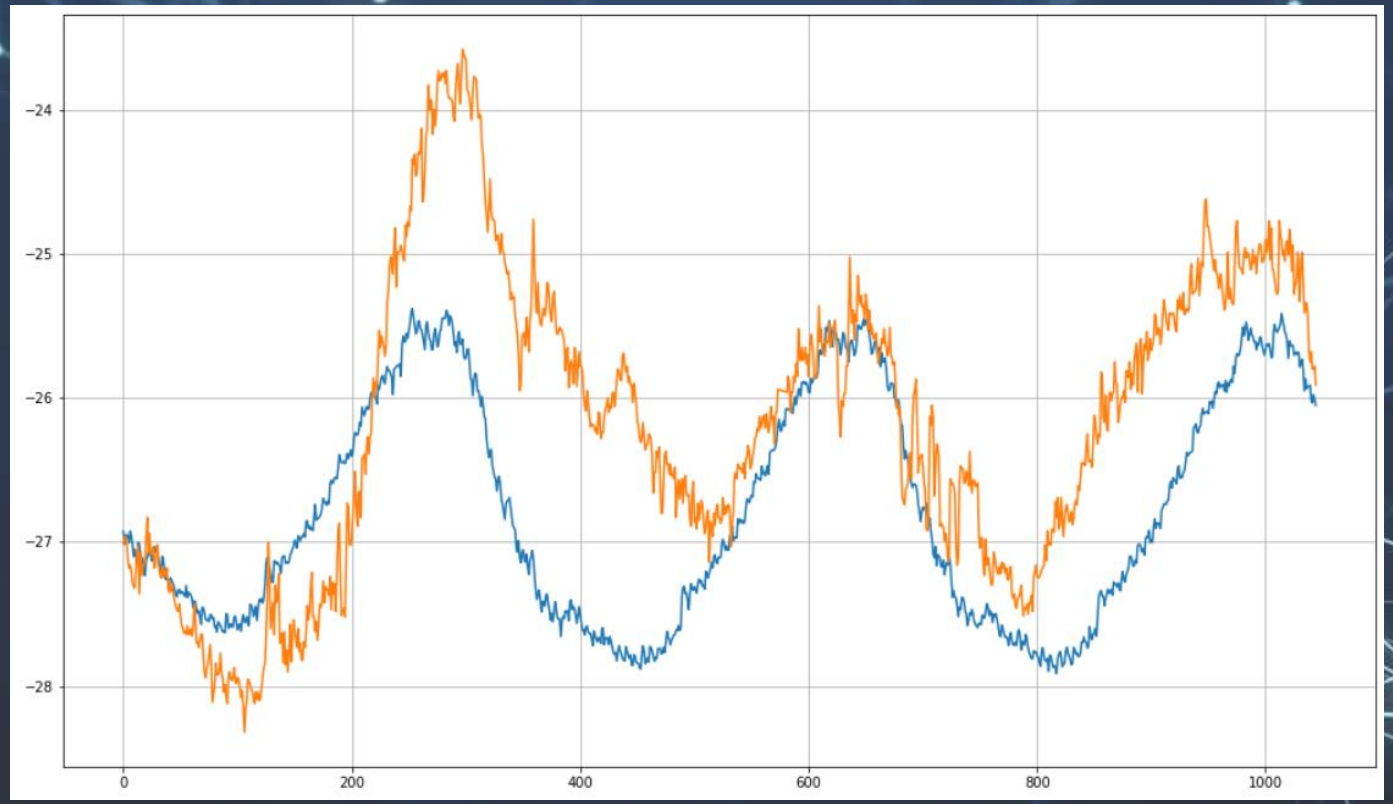
Dep. Variable:	y	**No. Observations:**	4178
Model:	SARIMAX(3, 1, 2)	**Log Likelihood**	3250.577
Date:	Tue, 12 Jan 2021	**AIC**	-6489.153
Time:	05:49:55	**BIC**	-6451.129
Sample:	0	**HQIC**	-6475.704
	- 4178		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	1.6207	0.035	46.228	0.000	1.552	1.689
ar.L2	-0.6577	0.032	-20.533	0.000	-0.721	-0.595
ar.L3	0.0339	0.016	2.111	0.035	0.002	0.065
ma.L1	-1.7567	0.034	-52.320	0.000	-1.823	-1.691
ma.L2	0.7643	0.033	22.965	0.000	0.699	0.830
sigma2	0.0123	0.000	71.578	0.000	0.012	0.013

Ljung-Box (L1) (Q):	0.00	**Jarque-Bera (JB):**	1628.35
Prob(Q):	0.99	**Prob(JB):**	0.00
Heteroskedasticity (H):	1.09	**Skew:**	0.31
Prob(H) (two-sided):	0.12	**Kurtosis:**	6.00

Petrigano

Тестирование



Petrigano

Предсказание

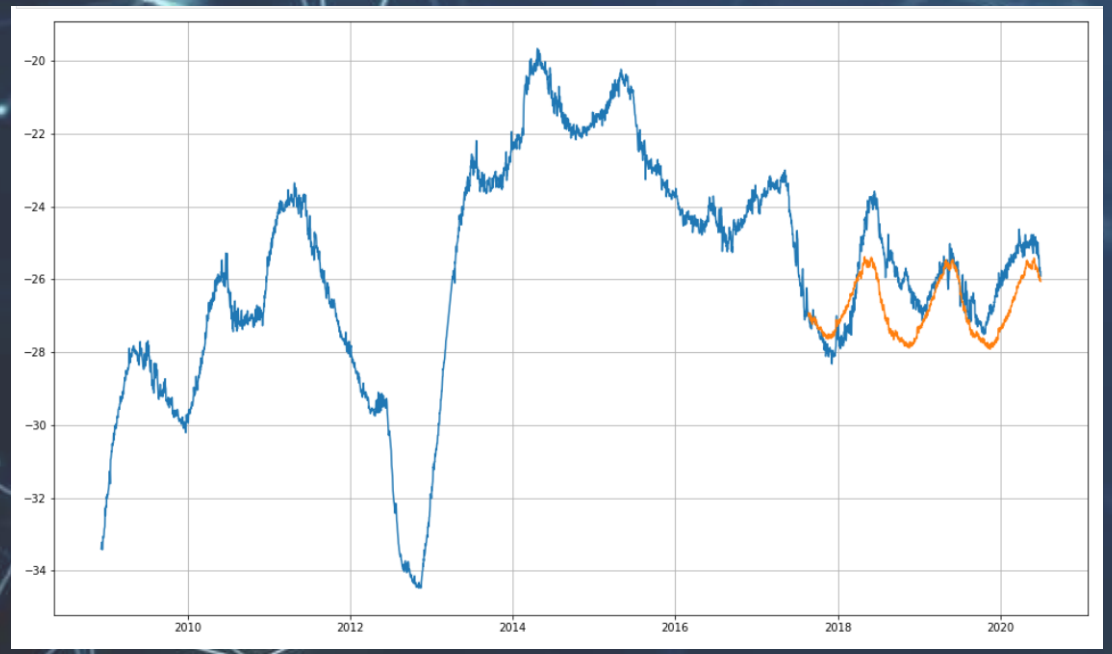
Сначала были последовательно восстановлены сезонности и временной ряд вернули на первоначальный уровень

Предсказание временного ряда

Предсказание данных

Petrigano

Результаты

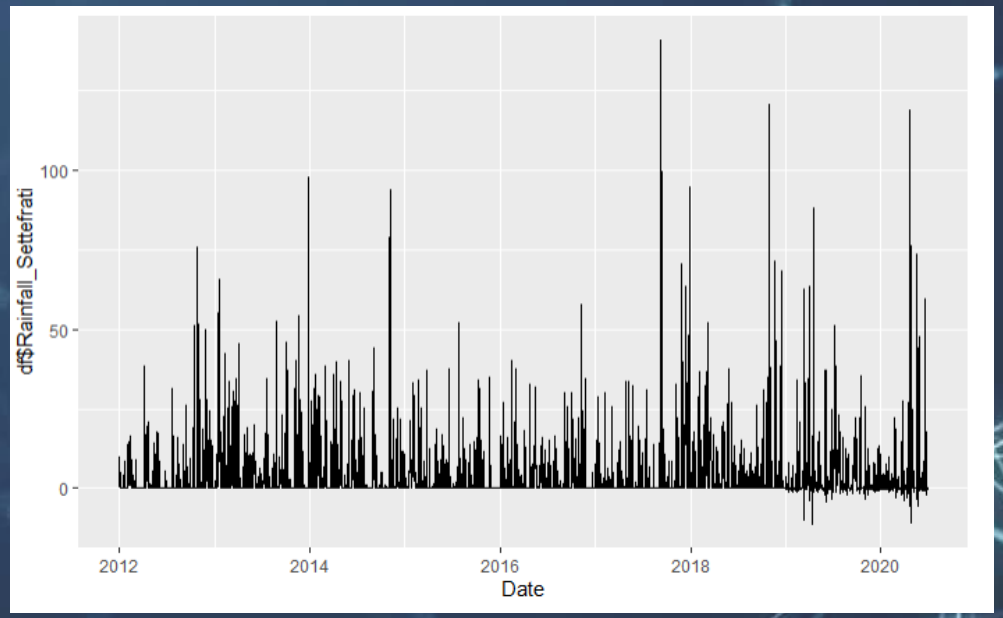


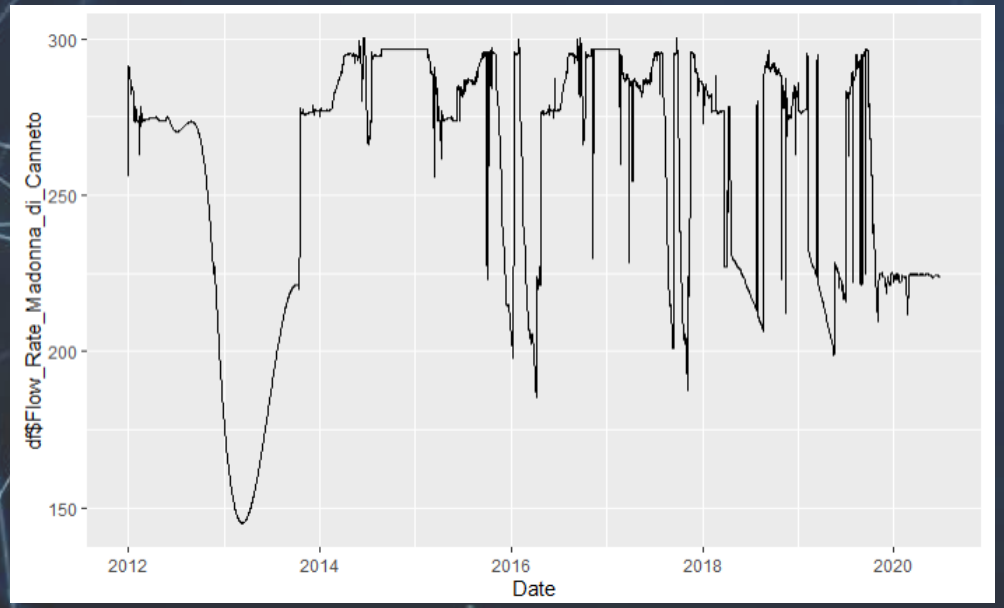
1. **Максимальный объем в марте, минимальный – в декабре**
2. **Максимальный расход воды (1м^3 / сек через поперечное сечение) в марте, минимальное – в декабре**
3. **Максимальная температура в августе, минимальная в декабре**

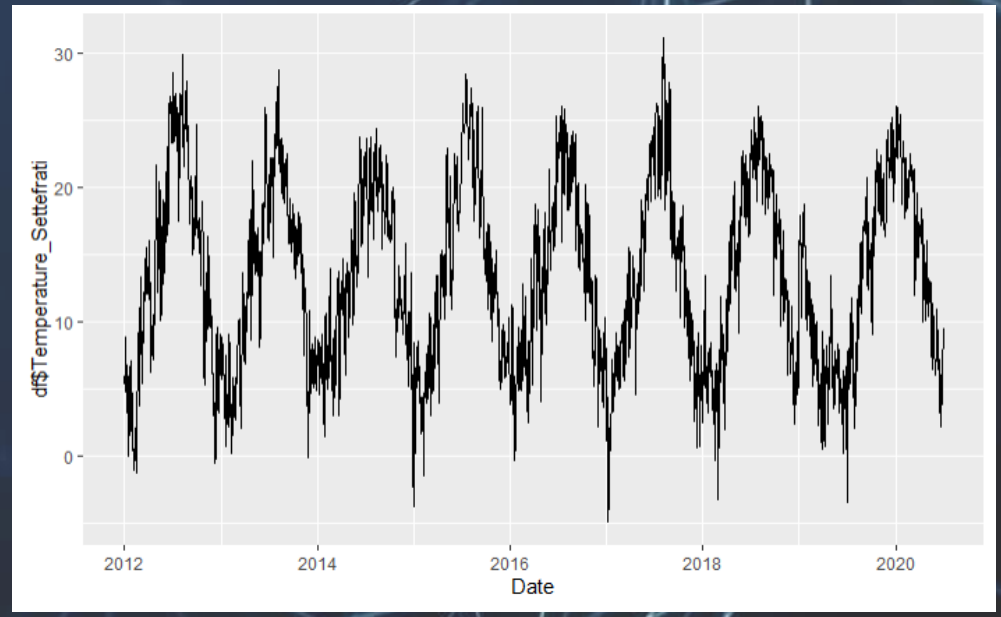
Предсказание данных

Water_Spring_Madonna_di_Canneto. Восстановление данных

Методом интерполяции восстанавливаем все нужные данные

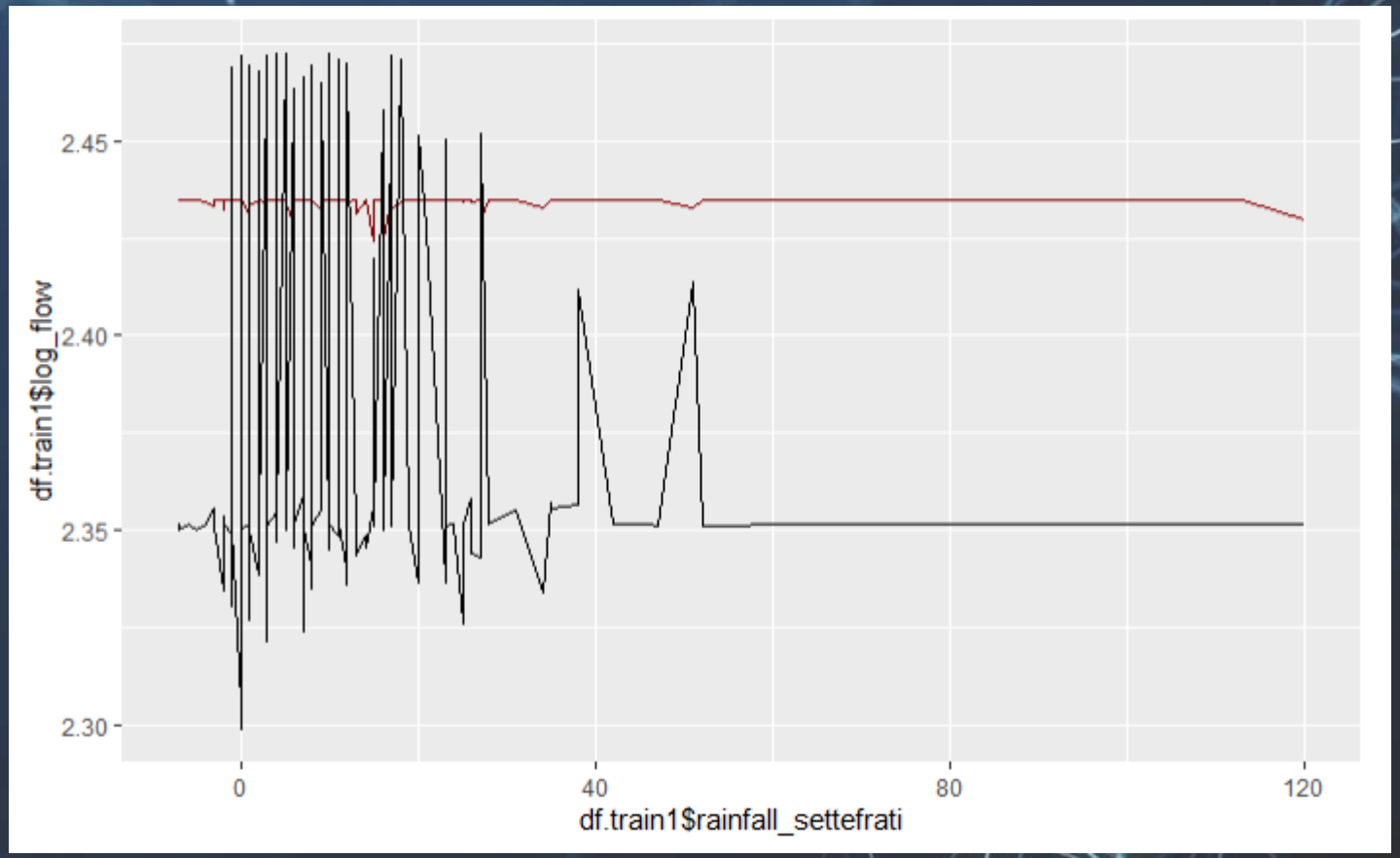






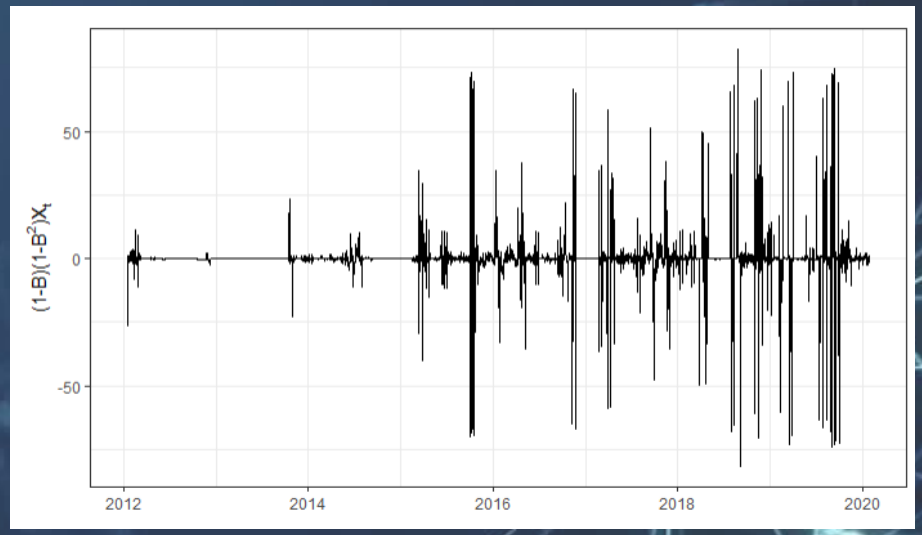
Water_Spring_Madonna_di_Canneto. Восстановление данных

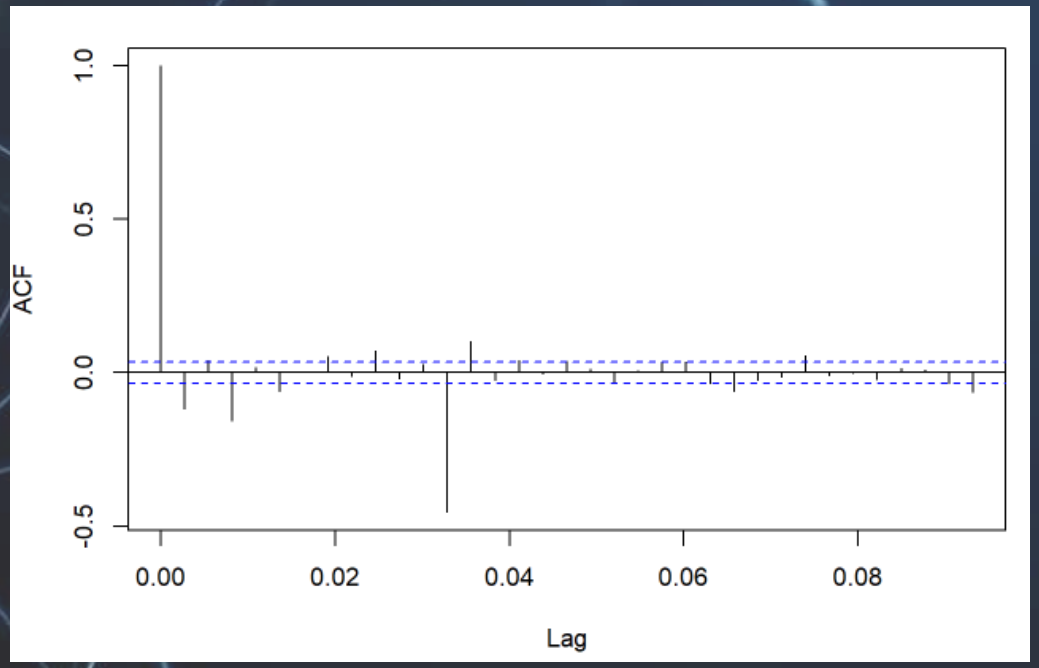
Методом линейной регрессии была предпринята попытка восстановления данных на основе имеющихся данных. Но полученная модель оказалась не точной, поэтому она была отклонена.

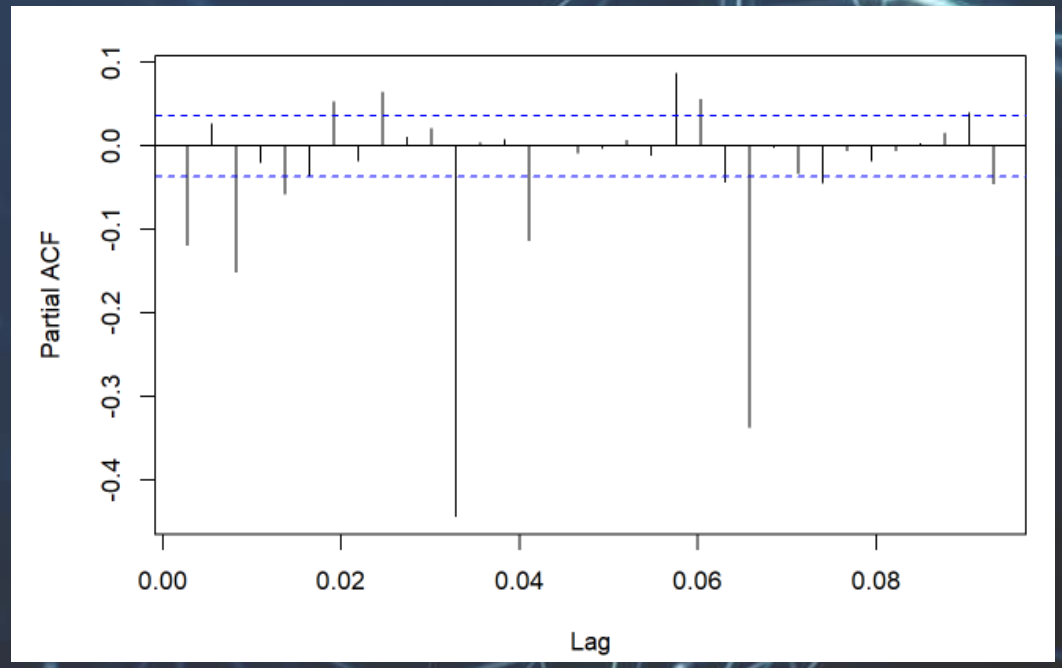


Water_Spring_Madonna_di_Canneto. Анализ данных

Проведено тестирование на стационарность данных (ADF test и KPSS) и их преобразование. Построены графики ACF и PACF, по ним были определены параметры для построения модели ARIMA.







Water_Spring_Madonna_di_Canneto. Анализ данных

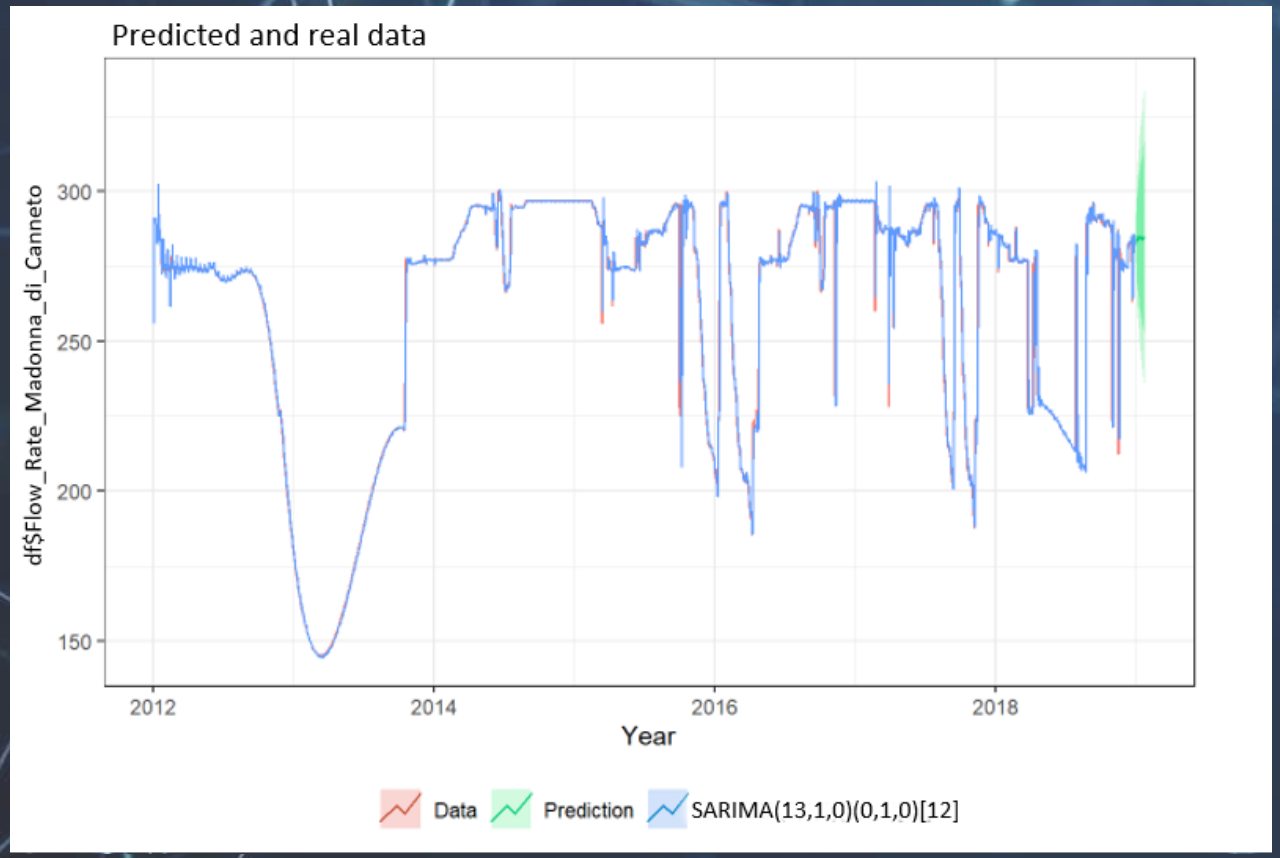
Были построены возможные модели ARIMA и проверены их коэффициенты, а так же построена их сводная таблица

p	d	q	P	D	Q	T	residuals	aic	model
1	1	1	1	1	1	12	n	20201.7741	ARIMA(1,1,1)(1,1,1)[12]
1	1	0	1	1	1	12	n	20230.7296	ARIMA(1,1,0)(1,1,1)[12]
1	1	12	1	1	0	12	n	20246.2248	ARIMA(1,1,12)(1,1,0)[12]
1	1	11	1	1	0	12	n	21120.9644	ARIMA(1,1,11)(1,1,0)[12]
1	1	10	1	1	0	12	y	21319.6211	ARIMA(1,1,10)(1,1,0)[12]
1	1	9	1	1	0	12	y	21321.0678	ARIMA(1,1,9)(1,1,0)[12]
1	1	8	1	1	0	12	y	21212.9521	ARIMA(1,1,8)(1,1,0)[12]
1	1	7	1	1	0	12	y	21319.8338	ARIMA(1,1,7)(1,1,0)[12]
1	1	6	1	1	0	12	y	21319.7493	ARIMA(1,1,6)(1,1,0)[12]
1	1	5	1	1	0	12	y	21318.3318	ARIMA(1,1,5)(1,1,0)[12]

Showing 1 to 10 of 255 entries

Previous 1 2 3 4 5 … 26 Next

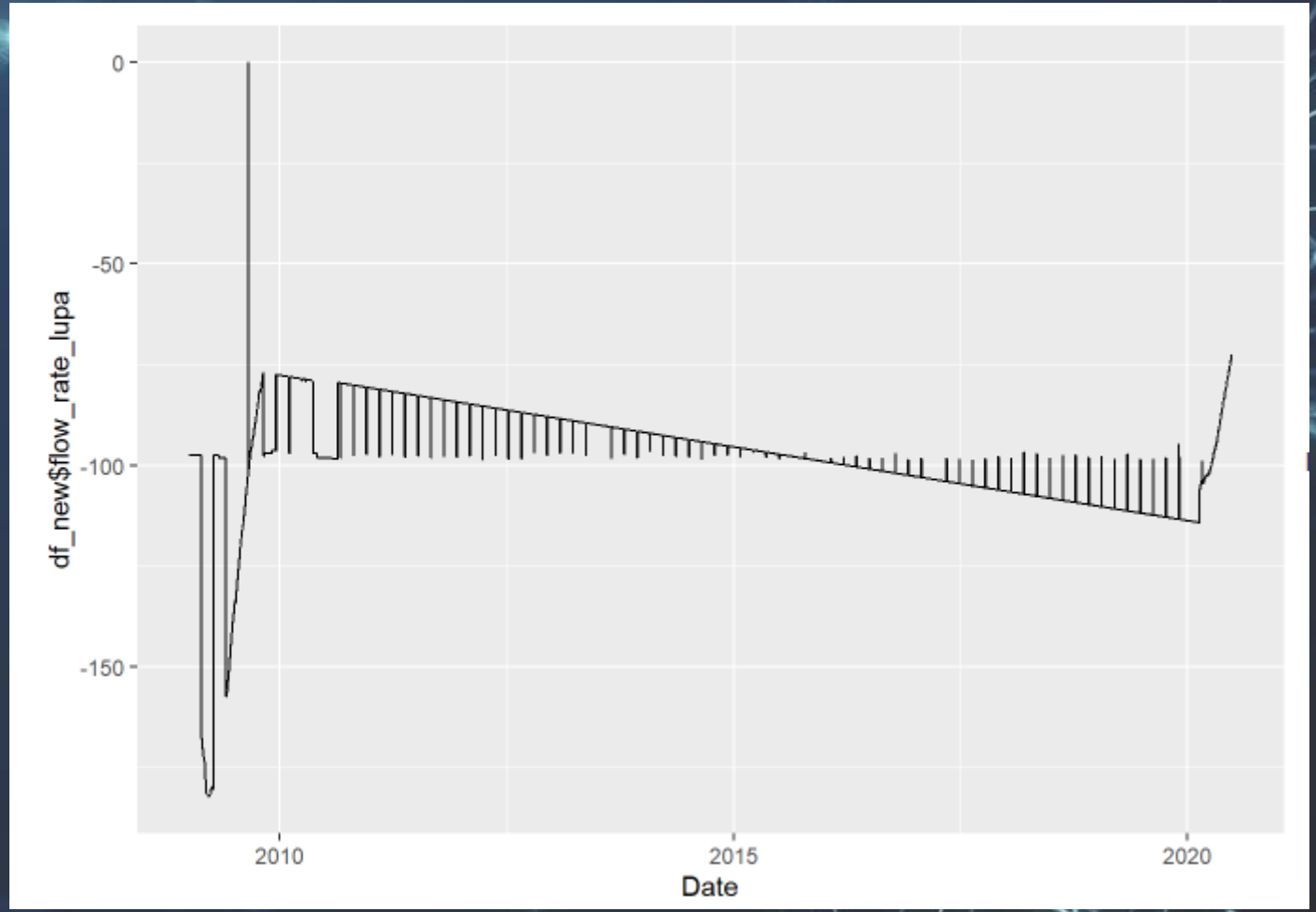
Water_Spring_Madonna_di_Canneto. Результаты



В качестве конечного результата была получена модель, которая позволяет прогнозировать данные о скорости потока в данном роднике.

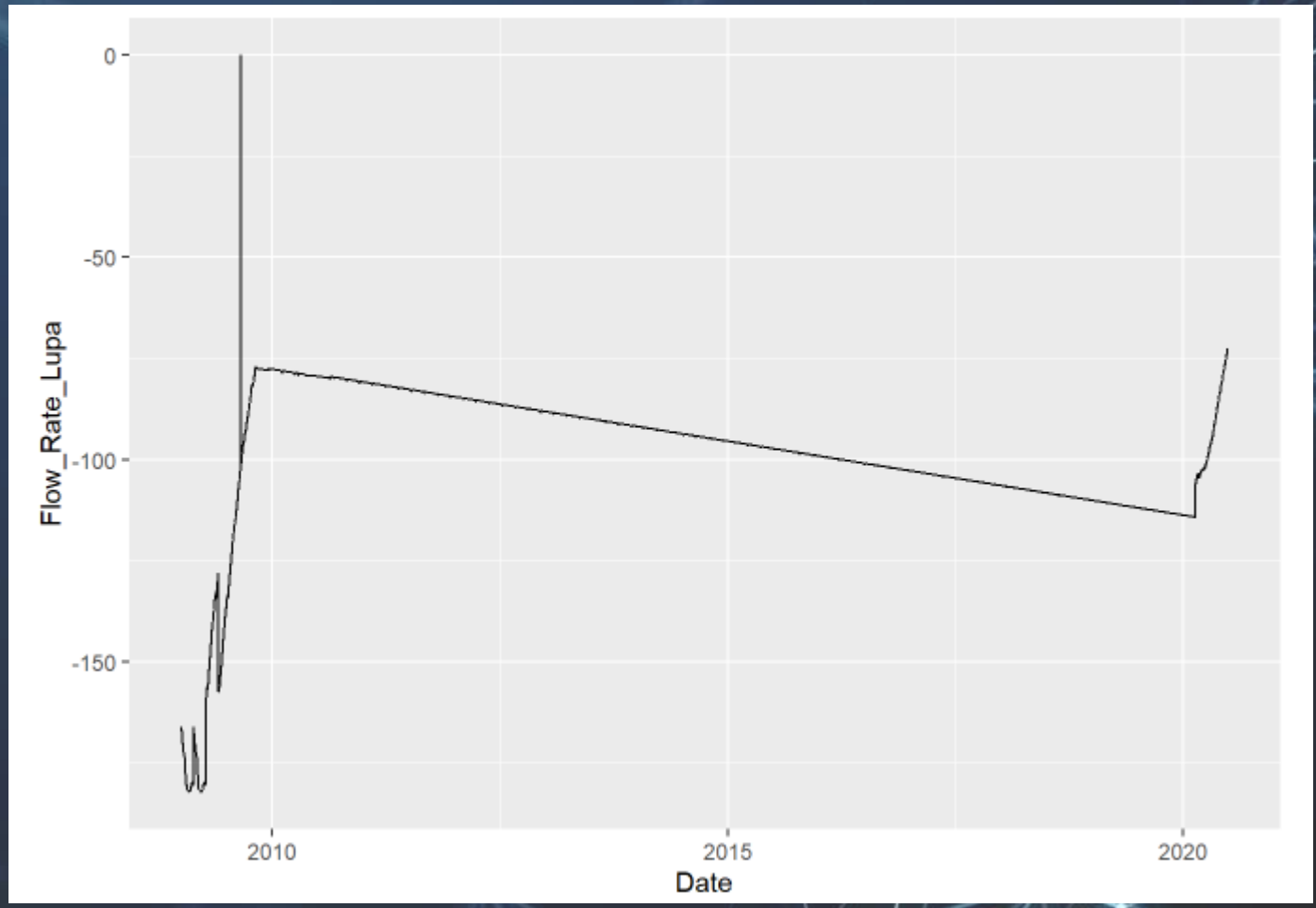
Water_Spring_Lupa. Восстановление данных

Методом линейной была предпринята попытка восстановления данных на основе имеющихся данных. Но полученная модель оказалась не точной, поэтому она была отклонена.



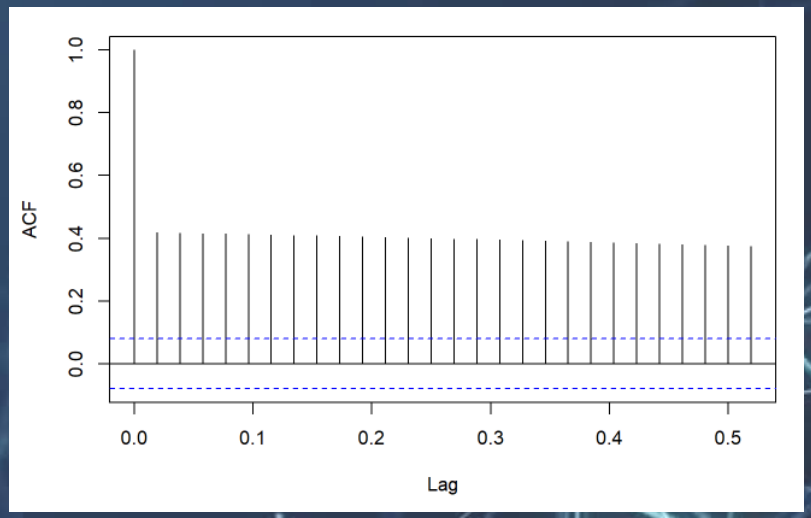
Water_Spring_Lupa. Восстановление данных

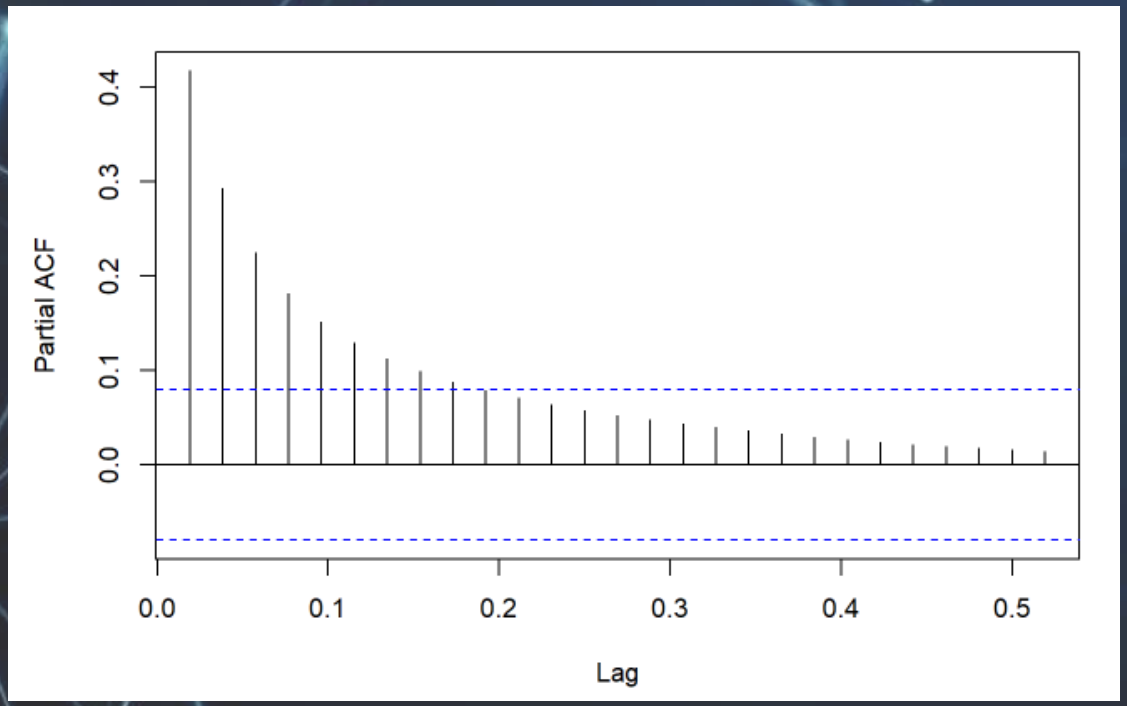
Методом интерполяции восстанавливаем все нужные данные.

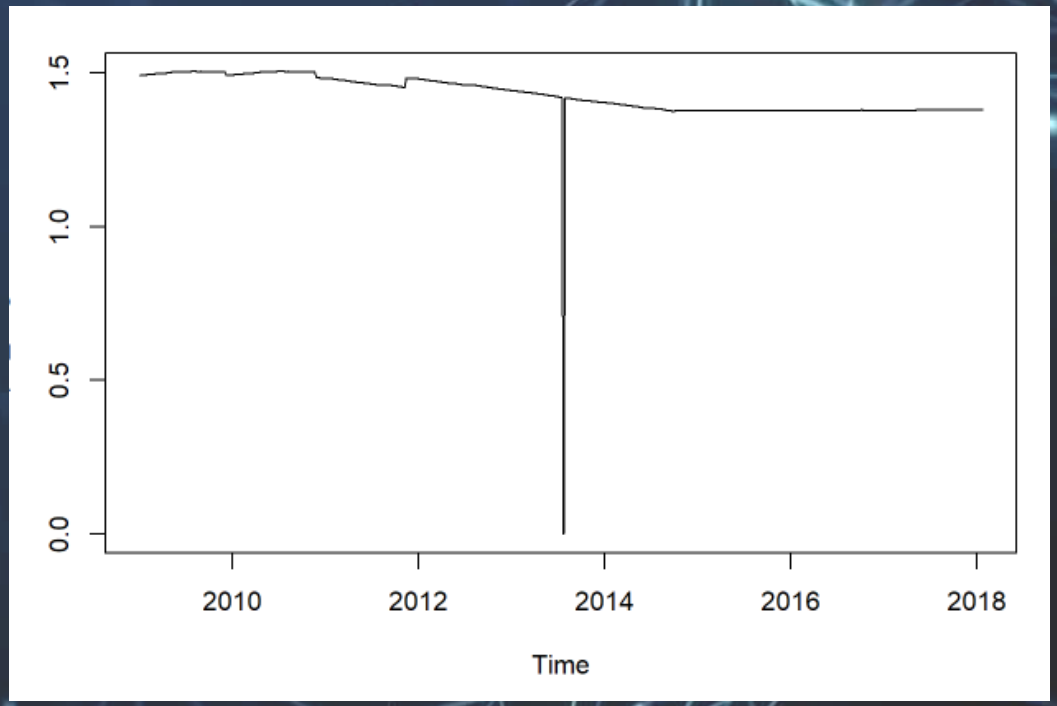


Water_Spring_Lupa. Анализ данных

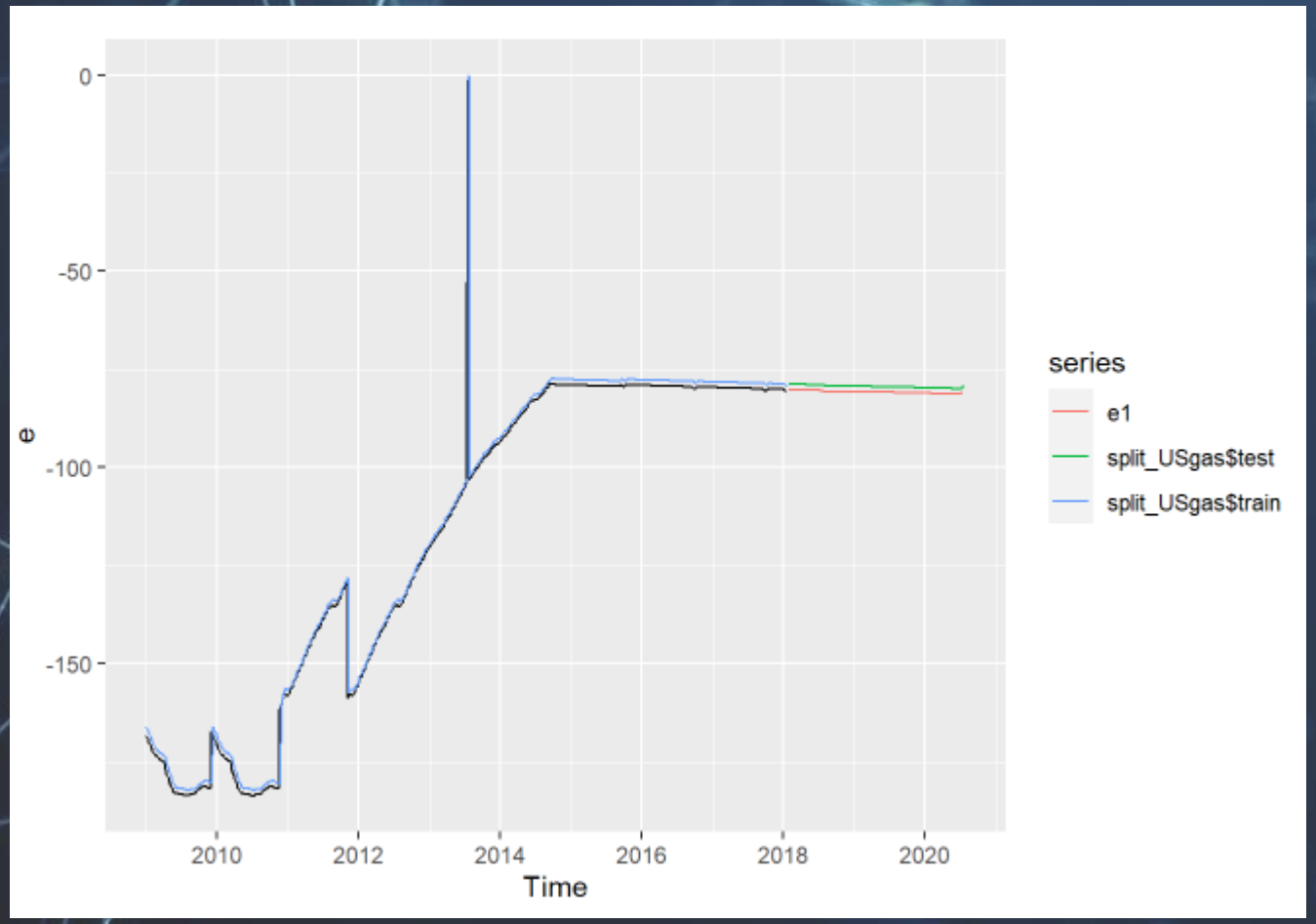
Проведено тестирование на стационарность данных (ADF test и KPSS) и их преобразование. Построены графики ACF и PACF, по ним были определены параметры для построения модели ARIMA.







Water_Spring_Lupa. Результаты



В качестве конечного результата была получена модель, которая позволяет прогнозировать данные о скорости потока в данном роднике.

Lake Bilancino. Подготовка данных

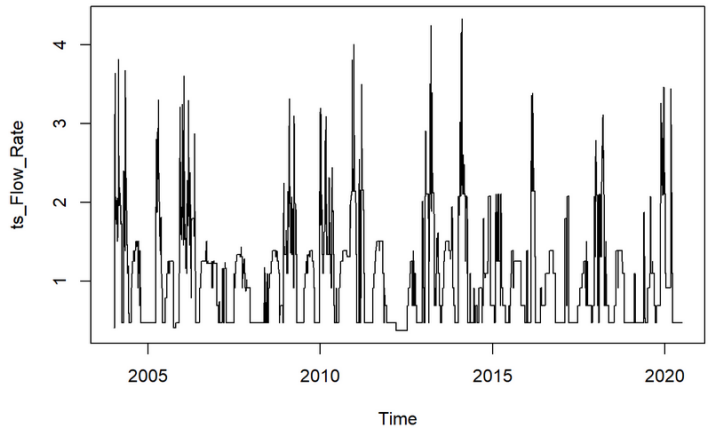
- **Проверяем, что данные расположены правильно (по дате)**
- **Первые записи для нас бесполезны – отрезаем их**
- **Проверяем стационарность, выполняем трансформации**

```
df$Date = strptime(df$Date, "%d/%m/%Y")
df$Date <- ymd(df$Date)

df <- df[order(df$Date), ]
df['Interval']=df$Date - shift(df$Date, n=1, fill=NA, type="lag")
days <- df$Interval

for(i in days)
{

  if(isTRUE(i > 1) || is.null(i))
  {
    print(i)
  }

}
```

```
df <- df[579:6603,]
summary(df)
```

```
##  Rainfall_S_Piero Rainfall_Mangona  Rainfall_S_Agata Rainfall_Cavallina
##  Min.   : 0.000   Min.   :  0.000   Min.   :  0.00   Min.   :  0.000
##  1st Qu.: 0.000   1st Qu.:  0.000   1st Qu.:  0.00   1st Qu.:  0.000
##  Median : 0.000   Median :  0.000   Median :  0.00   Median :  0.000
##  Mean   : 2.472   Mean   :  3.341   Mean   :  2.67   Mean   :  2.675
##  3rd Qu.: 0.800   3rd Qu.:  1.400   3rd Qu.:  1.20   3rd Qu.:  0.600
##  Max.   :80.600   Max.   :110.000   Max.   :120.20   Max.   :113.600
##  Rainfall_Le_Croci Temperature_Le_Croci   Lake_Level      Flow_Rate
##  Min.   : 0.00     Min.   :-5.35        Min.   :243.5   Min.   : 0.450
##  1st Qu.: 0.00     1st Qu.: 9.00        1st Qu.:247.9   1st Qu.: 0.600
##  Median : 0.00     Median :14.50        Median :250.2   Median : 1.500
##  Mean   : 3.13     Mean   :14.53        Mean   :249.6   Mean   : 2.778
##  3rd Qu.: 1.20     3rd Qu.:20.10        3rd Qu.:251.4   3rd Qu.: 3.000
##  Max.   :88.40     Max.   :34.00        Max.   :252.8   Max.   :74.650
##       Date
##  Min.   :2004-01-02
##  1st Qu.:2008-02-16
##  Median :2012-04-01
##  Mean   :2012-04-01
##  3rd Qu.:2016-05-16
##  Max.   :2020-06-30
```

```
ts_Flow_Rate <- log(ts(df$Flow_Rate+1, start=c(2004,2), frequency = 365.25))
plot(ts_Flow_Rate)
```



Lake Bilancino. Модели – как тренируем

Для построения моделей используется VAR. Было построено 4 модели, которые далее сравниваются на кросс-валидации.

```
y <- cbind(ts_Flow_Rate, ts_Lake_Level)
colnames(y)<-cbind("Flow_Rate", "Lake_Level")

x <- cbind(ts_Rainfall_Cavallina, ts_Rainfall_Le_Croci, ts_Rainfall_Mangona, ts_Rainfall_S
_Agata, ts_Rainfall_S_Piero, ts_Temperature_Le_Croci)
colnames(x)<-cbind("Rainfall_Cavallina", "Rainfall_Le_Croci", "Rainfall_Mangona", "Rainfal
l_S_Agata", "Rainfall_S_Piero", "Temperature_Le_Croci")
```

Now we need to define the p parameter for our model.

```
lagselect <- VARselect(y=y, type = "const", exogen = x)
lagselect$selection
```

AIC(n): 5 **HQ(n):** 3 **SC(n):** 3 **FPE(n):** 5

```
y_2 <- cbind(ts_Flow_Rate, ts_Lake_Level)
colnames(y_2)<-cbind("Flow_Rate", "Lake_Level")

x_2 <- cbind(shift(ts_Rainfall_Cavallina, n=2L), shift(ts_Rainfall_Le_Croci, n=2L), shift(ts_Ra
infall_Mangona, n=2L), shift(ts_Rainfall_S_Agata, n=2L), shift(ts_Rainfall_S_Piero, n=2L), shif
t(ts_Temperature_Le_Croci, n=2L))
colnames(x_2)<-cbind("Rainfall_Cavallina", "Rainfall_Le_Croci", "Rainfall_Mangona", "Rainfall_S
_Agata", "Rainfall_S_Piero", "Temperature_Le_Croci")
```

```
y_1 <- cbind(ts_Flow_Rate, ts_Lake_Level)
colnames(y_1)<-cbind("Flow_Rate", "Lake_Level")

x_1 <- cbind(shift(ts_Rainfall_Cavallina), shift(ts_Rainfall_Le_Croci), shift(ts_Rainfall_Mango
na), shift(ts_Rainfall_S_Agata), shift(ts_Rainfall_S_Piero), shift(ts_Temperature_Le_Croci))
colnames(x_1)<-cbind("Rainfall_Cavallina", "Rainfall_Le_Croci", "Rainfall_Mangona", "Rainfall_S
_Agata", "Rainfall_S_Piero", "Temperature_Le_Croci")
```

```
y_3 <- cbind(ts_Flow_Rate, ts_Lake_Level)
colnames(y_3)<-cbind("Flow_Rate", "Lake_Level")

x_3 <- cbind(ts_Rainfall_Cavallina, ts_Rainfall_Le_Croci, ts_Rainfall_Mangona, ts_Rainfall_S_Ag
ata, ts_Rainfall_S_Piero, ts_Temperature_Le_Croci, shift(ts_Rainfall_Cavallina), shift(ts_Rainf
all_Le_Croci), shift(ts_Rainfall_Mangona), shift(ts_Rainfall_S_Agata), shift(ts_Rainfall_S_Pier
o),shift(ts_Temperature_Le_Croci),shift(ts_Rainfall_Cavallina, n=2L), shift(ts_Rainfall_Le_Croc
i, n=2L), shift(ts_Rainfall_Mangona, n=2L), shift(ts_Rainfall_S_Agata, n=2L), shift(ts_Rainfall
_S_Piero, n=2L),shift(ts_Temperature_Le_Croci, n=2L))
colnames(x_3)<-cbind("Rainfall_Cavallina", "Rainfall_Le_Croci", "Rainfall_Mangona", "Rainfall_S
_Agata", "Rainfall_S_Piero", "Temperature_Le_Croci","Rainfall_Cavallina_1", "Rainfall_Le_Croci_
1", "Rainfall_Mangona_1", "Rainfall_S_Agata_1", "Rainfall_S_Piero_1","Temperature_Le_Croci_1","
Rainfall_Cavallina_2", "Rainfall_Le_Croci_2", "Rainfall_Mangona_2", "Rainfall_S_Agata_2", "Rain
fall_S_Piero_2","Temperature_Le_Croci_2")
```

Lake Bilancino. Модели – тестирование и кросс-валидация

Для построенных моделей проведено тестирование с прогнозированием на 2 года вперед. Проведена кросс-валидация для их сравнения.

```
border <- length(y[,1])-2*365
train_y <- y[1:border,]
train_x <- x[1:border,]
test_y <- y[(border+1):length(y[,1]),]
test_x <- x[(border+1):length(y[,1]),]
```

0.22631176560062 · 0.000252284348622253

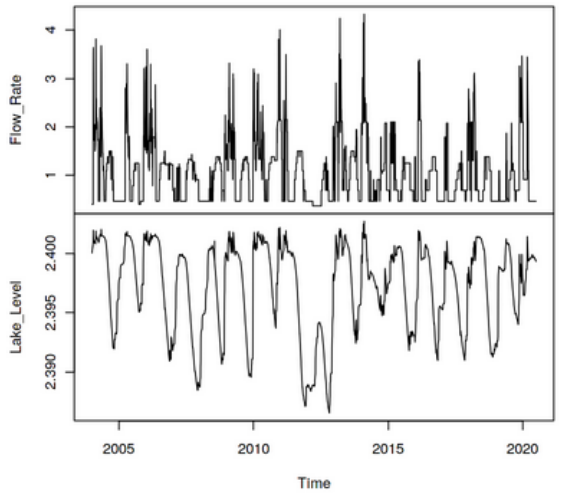
0.221260757818531 · 0.000224204822228104

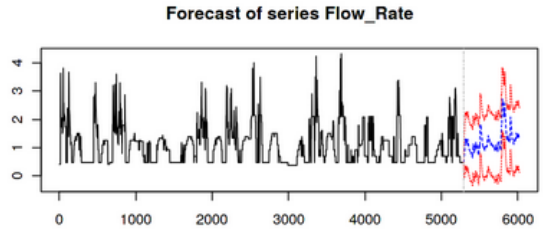
0.224814814896173 · 0.000229611915168508

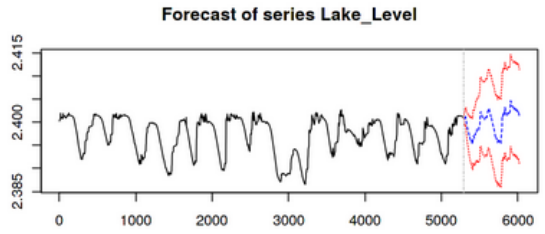
0.217157932997212 · 0.000210243470835219

```
cross_val <- function(x, y, p) {
  max <- length(y[,1])
  border <- max-2*365
  pred_length <- 3
  errors_Flow=0
  errors_Level=0
  measures = 0
  while (border+pred_length <= max) {
    train_y <- y[1:border,]
    train_x <- x[1:border,]
    test_y <- y[(border+1):(border+pred_length),]
    test_x <- x[(border+1):(border+pred_length),]
    test_Model <- VAR(train_y, p = p, type = "const",  exogen = train_x)
    pred_y <- predict(test_Model, n.ahead = pred_length, dumvar=test_x, ci=0.95)
    check<-pred_y$fcst$Flow_Rate
    errors_Flow <- sum((check[,1]-test_y[,1])^2)+errors_Flow
    check<-pred_y$fcst$Lake_Level
    errors_Level <- sum((check[,1]-test_y[,2])^2)+errors_Level
    measures=measures + 3
    border=border+pred_length
  }
  RMSE_Flow = sqrt(errors_Flow/measures)
  RMSE_Level = sqrt(errors_Level/measures)
  c(RMSE_Flow,RMSE_Level)
}
```

Lake Bilancino. Результаты







В качестве конечного результата была получена модель, которая позволяет прогнозировать данные о скорости потока воды и уровне воды в озере.

Также проведена её кросс-валидация с результатами прогнозирования, трансформированными к исходной форме, чтобы другие создатели моделей могли сравнивать ее с нашей.

RMSE_Flow_Rate = 2.26697309092565
RMSE_Level = 0.121072372651214