RAPORT Z REALIZACJI ETAPU PROJEKTU

(OSIĄGNIĘCIA KAMIENIA MILOWEGO)

NR ETAPU 1

W RAMACH PROGRAMU OPERACYJNEGO INTELIGENTNY ROZWÓJ

A. DANE PROJEKTU				
Numer umowy	**POIR.01.01.01-00-0134/17**			
Tytuł projektu	Predykcja wydajności sieci kanalizacyjno-burzowej w czasie rzeczywistym jako usługa SaaS oparta na danych pozyskanych metodami uczenia maszynowego.			
Okres realizacji etapu	**od**	**2017-09-01**	**do**	2018-06-30
Okres realizacji projektu: *(zgodnie z bieżącymi zapisami Umowy)*:	**od**	**2017-09-01**	**do**	**2018-08-31**

B. DANE BENEFICJENTA	
Nazwa Beneficjenta	**CARL Data Solutions pL sp z o. o.**
Imię i nazwisko osoby sporządzającej raport	**Piotr Stępiński**
Telefon kontaktowy	**505990555**
E-mail	**piotr@carlsolutions.com**

C. INFORMACJE DOTYCZĄCE KAMIENIA MILOWEGO

Etap nr: 1 realizowany w ramach badań przemysłowych / prac rozwojowych.[1]
Kamień milowy - nazwa: Model predykcyjny reakcji infrastruktury kanalizacyjnej na deszcze o skuteczności prognostycznej >80%, działający z wykorzystaniem topologii tejże sieci.
Poziom TRL[2] **osiągnięty po zakończeniu ww. Etapu: VIII**

Deklaracja Beneficjenta:	TAK	NIE
1. Czy etap zakończył się osiągnięciem kamienia milowego?	**X**	
2. Czy wszystkie zadania / prace w ramach etapu zostały zrealizowane?	**X**	
3. Czy Beneficjent wprowadził rekomendacje wskazane w ramach oceny poprzedniego raportu? *(jeśli dotyczy)*[3]		

[1] *niepotrzebne skreślić*
[2] *źródło: http://www.ncbir.pl/gfx/ncbir/pl/defaultopisy/1195/1/1/poziomy_gotowosci_technologicznej.pdf*
[3] *zaznaczyć wpisując „X" we właściwe pole*

W przypadku zaznaczenia opcji „TAK” należy opisać wdrożenie każdej rekomendacji. W przypadku zaznaczenia opcji „NIE” należy uzasadnić dlaczego nie wdrożono rekomendacji: Nie dotyczy

4. Podmiot odpowiedzialny za realizację etapu / prac (Beneficjent / nazwa Podwykonawcy)	
Beneficjent	
Sposób udokumentowania uzyskanych wyników[4]:	D raport opisujący wyniki
Dodatkowe sposoby udokumentowania wyników[5]	
Wskazać osiągnięty kamień milowy:	

5. Ewentualne odstępstwa od osiągnięcia zakładanego kamienia milowego (uzasadnić / podać przyczynę odstępstw oraz opisać skutki dla dalszej realizacji projektu/ czy wystąpiły ryzyka w etapie, o których mowa we wniosku o dofinansowanie).

W toku projektu wystąpiły ryzyka które zostały opisane jako wynikające z charakteru danych zbyt rzadkie lub zbyt słabe korelacje.
Ryzyko sezonowości i silnych trendów zostało wykorzystywane jako dodatkowa informacja w budowanych modelach predykcyjnych i zamienione na atut który zniwelował niemożność wykorzystywania topologii.

* automatyczne wykrywanie topologii w trakcie pracy badawczej okazało się trudne/ niemożliwe do wykonania gdyż w trakcie eksploracji danych odnaleziono przypadki dla których różne punkty sieci wydawaly sie skorelowane a tymczasem nie wynikało to z danych GIS oraz że matematycznie nie jest potwierdzona korelacja która jest znana z danych GIS. Występowanie takich przypadkow praktycznie wykluczyło dalsza sciezke badania polegająca na próby algorytmicznego odtwarzania topologii z danych. Występowanie tych zależności także uniemożliwia wykorzystanie topologii do predykcji czy poprawienia skuteczności wykrywania anomalii.
* znaleziono model predykcyjny o zakładanej skuteczności który nie uwzględnia topologii
* tenże model / modele umożliwia także wykrywanie anomalii na zakładanym poziomie

[4] *Należy podać symbol i opis sposobu potwierdzenia przeprowadzonych prac i uzyskanych wyników: D - dokumentacja (np. dokumentacja techniczna, opracowanie założeń do prototypu, linii technologicznej, procesu) - symbol, numer, nazwa, data itp.; W - udokumentowane wyniki pomiarów; R - raporty (raporty cząstkowe opisujące przeprowadzone prace) - symbol, nazwa; data Z - zgłoszenie o certyfikację lub uznanie zgodności z normą - numer zgłoszenia, data zgłoszenia lub uznania zgodności z normą; ZP - zgłoszenie patentowe, patent - numer; data zgłoszenia, , C - uzyskane certyfikaty - numer; data P - publikacja, prezentacja, wydanie książkowe; (należy wskazać datę publikacji, autor i źródło), I - inne - jeśli wymienione kategorie nie wypełniają sposobu potwierdzenia rezultatów prac, należy wpisać literę I oraz podać krótki opis. W przypadku pozyskania informacji od opiekuna merytorycznego projektu w IP o konieczności uzupełnienia Raportu o dokumentację potwierdzającą osiągnięte rezultaty należy je przekazać tylko w formie elektronicznej bezpośrednio do opiekuna merytorycznego projektu w IP - w formacie pdf.*

[5] *W przypadku pozyskania informacji od opiekuna merytorycznego projektu w IP o konieczności uzupełnienia Raportu dopuszczalne jest również dodatkowe przekazanie plików z filmami (mov, avi, mp4, mkv, itp.), prezentacjami (np. PowerPoint, Prezi itp.) oraz plikamigraficznymi (jpg, tiff, png , itp.). Jeśli zaistnieje potrzeba ww. pliki należy przekazać bezpośrednio do opiekuna merytorycznego projektu.*

D. STOPIEŃ REALIZACJI WYDATKÓW W RAMACH ETAPU		
1. Planowane koszty realizacji etapu i poniesione/rzeczywiste koszty realizacji etapu	Koszty realizacji etapu planowane we wniosku o dofinansowanie w zł	Rzeczywiste koszty realizacji etapu
	932876.10	932876.10
W przypadku wystąpienia rozbieżności należy uzasadnić:		

E. CELOWOŚĆ DALSZEJ REALIZACJI PROJEKTU		
1. Czy zasadna jest kontynuacja realizacji projektu?	**TAK**	**NIE**
	X	
(W przypadku odpowiedzi „NIE” należy uzasadnić konieczność zaniechania realizacji projektu)		
2. Ewentualne działania naprawcze jakie należy podjąć w kolejnych etapach projektu, w przypadku gdy zostały zidentyfikowane odstępstwa w pkt. C.5. *(Syntetycznie opisać/uzasadnić konieczne do wprowadzenia zmiany w projekcie i ich wpływ na osiągnięcie rezultatów projektu - dotyczy tylko przypadku nieosiągnięcia zakładanych efektów/ rezultatów etapu)*		
Fakt że uzyskano wystarczająco skuteczne modele predykcyjne dla przepływu bez uwzględnienia topologii sieci oraz że model wykrywania anomalii ma także zakładaną skuteczność, powoduje że wykrywanie / uwzględnienie topologii staje się neutralne dla powodzenia projektu oraz nie są potrzebne żadne działania naprawcze		

F. DZIAŁANIA INFORMACYJNO-PROMOCYJNE W RAMACH REALIZOWANEGO PROJEKTU[6]

[6] Zasady Działań informacyjno - promocyjnych zostały zawarte m.in. w następujących dokumentach „Podręczniku wnioskodawcy i beneficjenta programów polityki spójności 2014-2020 w zakresie informacji i promocji” opublikowanym na stronie internetowej www.poir.gov.pl oraz w Wytycznych w zakresie promocji projektów finansowanych ze środków Narodowego Centrum Badań i Rozwoju, zamieszczonych na stronie www.ncbr.gov.pl

w ramach projektu prowadzone są działania informacyjno - promocyjne zgodnie zapisami § umowy o dofinansowanie dot. tych działań?	TAK	NIE
	X	

(W przypadku odpowiedzi „TAK" należy opisać, jakie działania są realizowane w ramach obowiązków informacyjno - promocyjnych projektu. W przypadku odpowiedzi „NIE", należy opisać dlaczego Beneficjent nie wypełnia tych obowiązków oraz jakie i kiedy zostaną wprowadzone środki zaradcze w tym zakresie.)

- informacja na stronie internetowej firmy
- informacja w biurze firmy

G. SZCZEGÓŁOWY OPIS ZREALIZOWANYCH PRAC ORAZ UZYSKANYCH WYNIKÓW W RAMACH ETAPU

(nie więcej niż 10 stron formatu A4 obejmujących opis zrealizowanych prac oraz osiągniętych rezultatów w okresie sprawozdawczym ze szczególnym uwzględnieniem metodologii oraz uzyskanych wyników przeprowadzonych badań przemysłowych lub prac rozwojowych, wytworzonych prototypów lub linii pilotażowych. W opisie rezultaty mogą być przedstawione w formie rysunków, schematów, wykresów, tabel, zdjęć. Opis powinien zawierać najistotniejsze informacje o uzyskanych wynikach - raport z kamienia milowego podlega ocenie, od której uzależniona jest kontynuacja finansowania projektu przez IP.)

1. Metodologia

- Jako metrykę do porównania / oceny uzyskanych w tym etapie modeli predykcyjnych przyjęto 90 percentyl MAE - Mean Absolute Error:

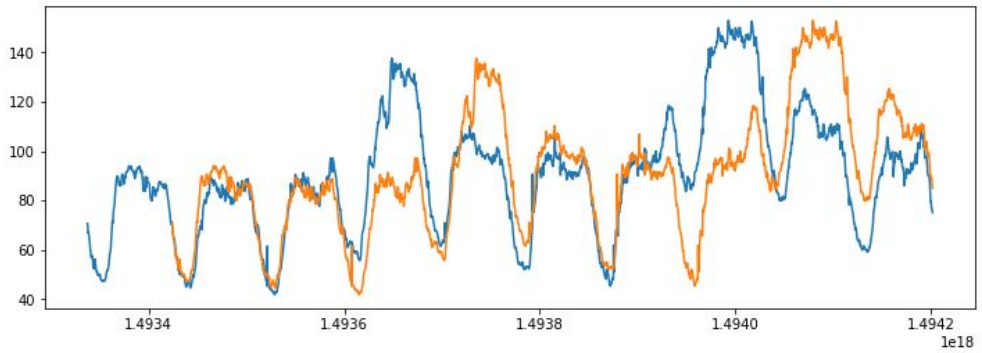
```
class PredictionModel:

    def fit(self, X, Y):
        pass

    def predict(self, X):
        pass


def mae(y_hat, y):
    """
    Calculate Mean Absolute Error
    """
    return np.sum(np.absolute(y_hat-y), axis=1)/y.shape[0]


def evaluate_model(model):
    """
    Evaluate model on all days starting from split_day.
    Returns 90th percentile error as model score
    """
    model.fit(X_train, Y_train)
    costs = mae(model.predict(X_test), Y_test)
    return np.percentile(costs, 90), costs
```

- Założono że kluczowe w projekcie jest zweryfikowanie hipotezy 3 tego etapu. Przyjęto że należy ten etap zaplanować w odwrotnej kolejności niż opisana pierwotnie.
    a. znalezienie bazowego / naiwnego modelu predykcyjnego przepływu który byłby punktem odniesienia dla pozostałych modeli

In [5]:
```
class ConstantMeanModel(PredictionModel):

    def __init__(self):
        self.mu = 0

    def fit(self, X, y):
        self.mu = np.mean(y)

    def predict(self, X):
        return np.ones(X.shape) * self.mu

score, costs = evaluate_model(ConstantMeanModel())
print('ConstantMeanModel score: {:.2f}'.format(score))
```

```
ConstantMeanModel score: 14.97
```

In [10]:
```
plot_days('2017-04-28', '2017-05-8')
```



Na powyższym obrazku pokazano fragment analizy przykładowego modelu który do predykcji stosuje wartość średnią poprzednich wartości. Model ten ma score 14.97.

Założyliśmy że aby ocenić przydatność dodatkowych informacji do realizacji celu projektu należy uzyskać modele które będą lepsze od bazowego.

b. znalezienie najlepszego możliwego modelu predykcyjnego przy założeniu że na wejściu mamy tylko informację o jednym przepływie

ConstMeanModel

In [4]:
```
class ConstantMeanModel(PredictionModel):

    def __init__(self):
        self.mu = 0

    def fit(self, xs):
        self.mu = np.mean(xs)

    def predict(self, day):
        return np.ones(12*24) * self.mu


score, costs = evaluate_model(ConstantMeanModel(), pd.Timestamp('2016-11-11'))
print('ConstantMeanModel score: {:.2f}'.format(score))
```

```
ConstantMeanModel score: 18.86
```

Previous Day Model

Uses values from last day

In [5]:
```
class LastDayModel(PredictionModel):

    def fit(self, xs):
        self.y = xs.values[-288:]

    def predict(self, day):
        return self.y


score, costs = evaluate_model(LastDayModel(), pd.Timestamp('2016-11-11'))
print('LastDayModel score: {:.2f}'.format(score))
```

```
LastDayModel score: 11.99
```

Model for single day. Easy case

In [6]:
```
evaluate_day(LastDayModel(), pd.Timestamp('2016-11-11'))
```

Out[6]:
```
4.7907965798611123
```

And when next day is kind of outlier

In [7]:
```
evaluate_day(LastDayModel(), pd.Timestamp('2017-05-01'))
```

Out[7]:
```
16.756769493055554
```

Daily Pattern model

Create pattern of daily usage based on historical data. Use this pattern to predict next values

(This can take up to 10 minutes to calculate)

In [8]:
```
class DailyPatternModel(PredictionModel):

    def fit(self, xs):
        df = flow.to_frame().reset_index()
        self.daily_pattern = df.groupby(by=[df.time.map(lambda x : (x.hour, x.minute))]).flow.mean().values

    def predict(self, day):
        return self.daily_pattern


score, costs = evaluate_model(DailyPatternModel(), pd.Timestamp('2016-11-11'))
print('DailyPatternModel score: {:.2f}'.format(score))
```

```
DailyPatternModel score: 9.61
```

Daily Pattern Median Model

Calculate median value for each time. Use it as a prediction for the next day.

In [14]:
```
class DayMedianModel(PredictionModel):

    def fit(self, xs):
        df = flow.to_frame().reset_index()
        self.daily_pattern = df.groupby(by=[df.time.map(lambda x : (x.hour, x.minute))]).flow.median().values

    def predict(self, day):
        return self.daily_pattern


score, costs = evaluate_model(DayMedianModel(), pd.Timestamp('2016-11-11'))
print('DayModel score: {:.2f}'.format(score))
```

```
DayModel score: 9.73
```

Na powyższym listingu pokazano fragment kodu w którym porównujemy różne modele dla tego samego zestawu danych. Widzimy tu że najlepszym modelem jest ten który wykorzystuje wzorzec dzienny i medianę. Jest on wystarczająco skuteczny żeby uznać go za spełniający cel projektu.

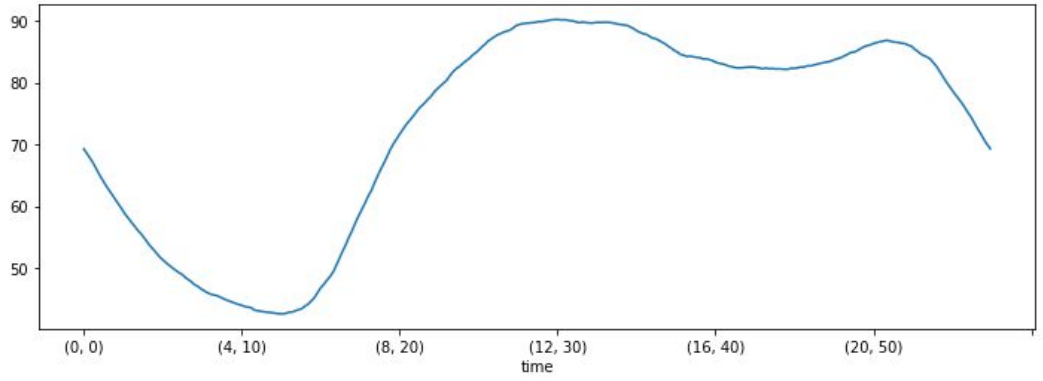
c. poszukiwanie najbardziej skutecznego modelu dla jednego przepływu uwzględniającego informację o deszczu

Założenie: Jeżeli nie uda się znaleźć modelu o lepszym score niż najlepszy z punktu b oznacza to że informacja o deszczu nie poprawia predykcji.

- Wykorzystanie metod używanych w analizie Inflow and Infiltration. Próba rozbicia predykcji dla “dni deszczowych” jako połączenia modelu wzorca dla dni suchych oraz dodatkowego przepływu wynikającego z opadu.
  Poniżej przedstawiono wyniki eksploracji wzorca dni suchych:

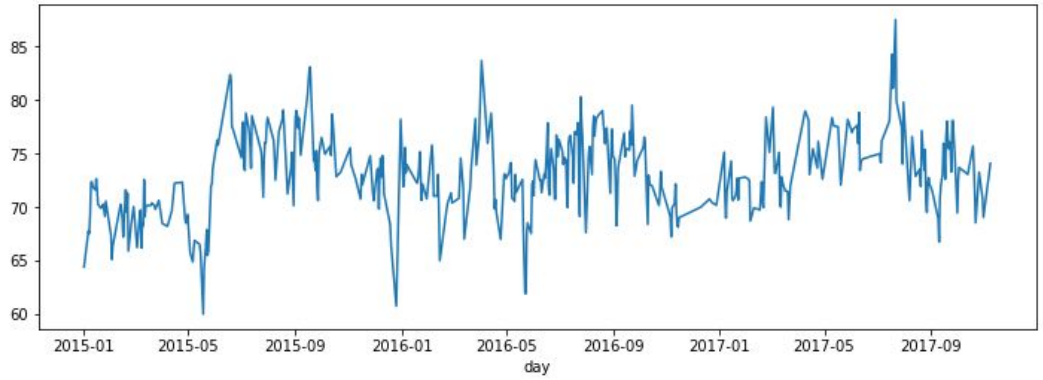
Mean DWP

```
In [18]: df_dry_days = data_frame[data_frame.day.isin(dry_days)]
         df = df_dry_days.reset_index()
         daily_pattern = df.groupby(by=[df.time.map(lambda x : (x.hour, x.minute))]).flow1.mean()
         daily_pattern.plot()
         plt.show()
```



How the DWP changes during the year

```
In [19]: df = df_dry_days.groupby(by=['day']).flow1.mean()
         df.plot()
         plt.show()
```



- Porównanie kilku modeli wykorzystujących informację o deszczu - regresja liniowa, drzewa decyzyjne, XGBoost

Linear regression

As a baseline lets try Linear Model

In [17]:
```
from sklearn.linear_model import LinearRegression

class LinearModel:

    def __init__(self, rain_window_size=1):
        self.rain_window_size = rain_window_size
        self.clf = LinearRegression()

    def fit(self, flow, rain):
        X, y = encode_features(flow, rain)
        self.clf.fit(X.values, y.values)

    def predict(self, day, rain, last_flow):
        base_features = prepare_prediction_features(day, rain).values
        predictions = []
        flow = last_flow
        for row in base_features:
            feature = np.array([[row[0], flow, row[1]]])
            pred = self.clf.predict(feature)[0]
            predictions.append(pred)
            flow = pred
        return np.array(predictions)

start_time = time.time()
model = LinearModel()
score, costs = evaluate_model(model, flow_rain.flow, flow_rain.rainfall, pd.Timestamp('2017-01-01'))
print('LinearModel score: {:.2f}%'.format(score))
print("Calculated in {:.3f} seconds".format(time.time() - start_time))
print('Model coef: {}, intercept: {}'.format(model.clf.coef_, model.clf.intercept_))
```

```
LinearModel score: 21.71%
Calculated in 304.371 seconds
Model coef: [ -9.37087226e-05   9.96541233e-01   7.65309783e-02], intercept: 0.41214071862701473
```

Decision Tree Regressor

First non linear model. Should improve on linear model

In [10]:
```
from sklearn import tree

class DTModel(LinearModel):

    def __init__(self):
        self.clf = tree.DecisionTreeRegressor()


start_time = time.time()
model = DTModel()
score, costs = evaluate_model(model, flow_rain.flow, flow_rain.rainfall, pd.Timestamp('2017-01-01'))
print('DTModel 2h score: {:.2f}%'.format(score))
print("Calculated in {:.3f} seconds".format(time.time() - start_time))
model.clf.feature_importances_
```

```
DTModel 2h score: 18.10%
Calculated in 567.181 seconds
```

XGBoost

In [27]:
```
import xgboost as xg

class XGBoostModel(LinearModel):

    def __init__(self, rain_window_size=2):
        self.rain_window_size = rain_window_size
        self.clf = xg.XGBRegressor()


start_time = time.time()
score, costs = evaluate_model(XGBoostModel(2), flow_rain.flow, flow_rain.rainfall, pd.Timestamp('2017-01-01'))
print('XGBoostModel 2h score: {:.2f}'.format(score))
print("Calculated in {:.3f} seconds".format(time.time() - start_time))
```

```
XGBoostModel 2h score: 17.36
Calculated in 871.612 seconds
```

Najskuteczniejszym modelem dla badanego datasetu był wykorzystujący regresję XGBoost.

2. Hipotezy i wnioski

W trakcie pracy badawczej potwierdzono hipotezę 3: *Jest możliwe zbudowanie modelu predykcyjnego reakcji infrastruktury kanalizacyjnej na deszcze o skuteczności prognostycznej >80%.*

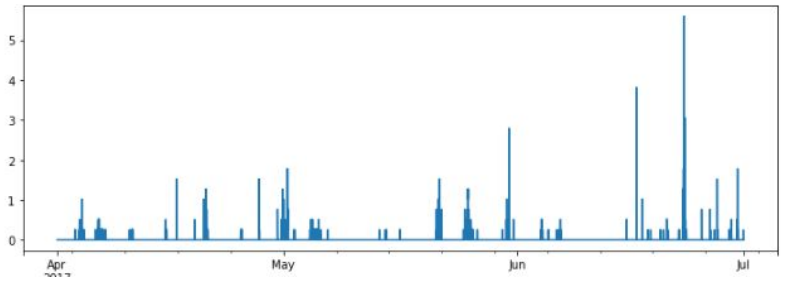
Nawet podstawowy model wykorzystujący wzorzec dzienny uzyskiwał założoną skuteczność.

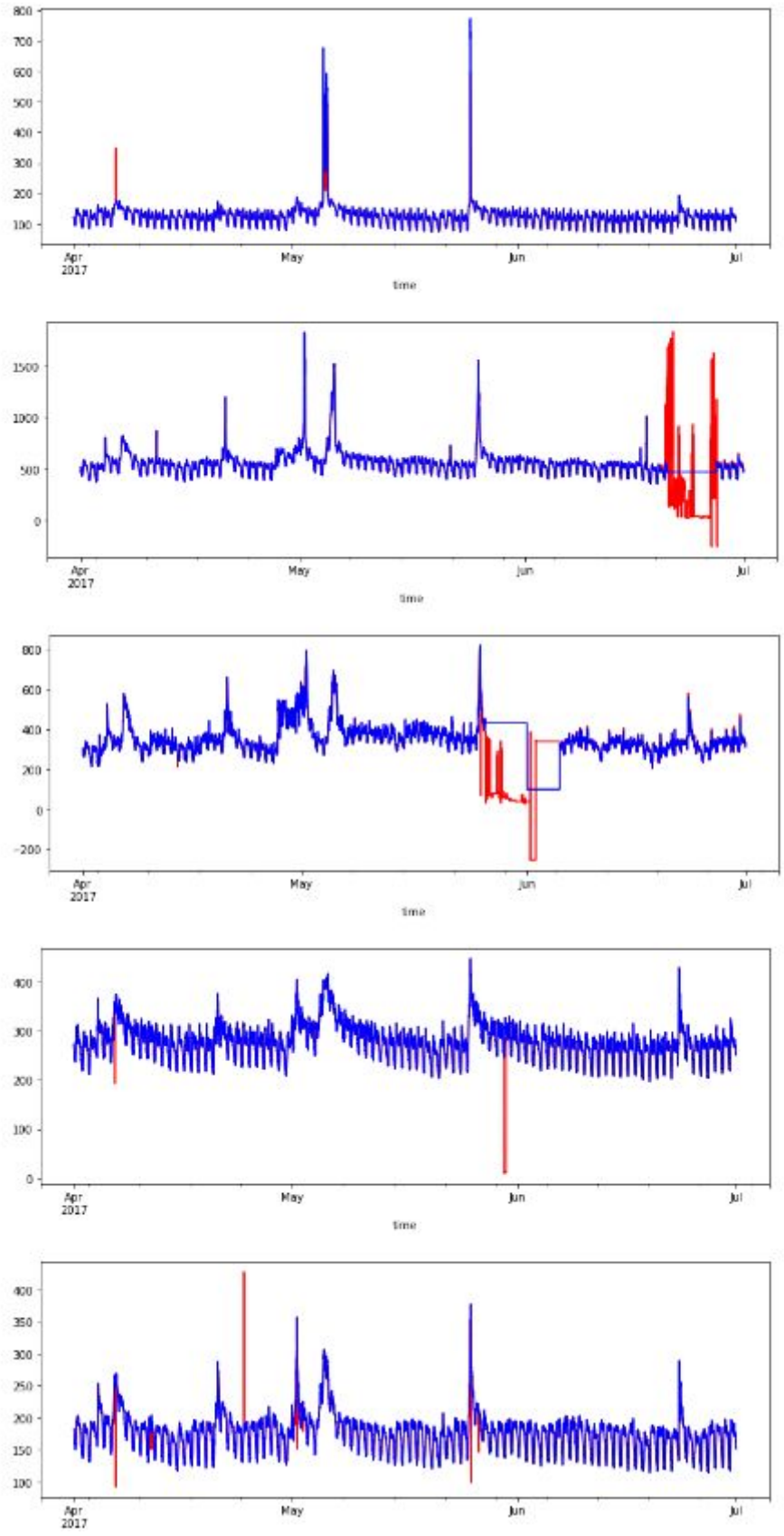
Co do hipotezy drugiej związanej ze zwiększeniem skuteczności wykrywania anomalii o 10% przy wykorzystaniu topologii to dla badanych w trakcie etapu danych nie potwierdzono tej hipotezy.

Na etapie eksploracji danych wykazano że istnieją kanały dla których z topologii wynikałaby korelacja, natomiast ich pobudzenie w wyniku deszczu jest niezależne od tej korelacji. Uniemożliwiło to modelowanie z uwzględnieniem tejże korelacji.

Poniżej pokazano przykładowy listing który to obrazuje:

```
In [4]: df = data_frame['2017-04-01': '2017-06-30']

        df.rainfall1.plot()
        plt.show()

        df.flow1.plot(color='r')
        df.flow1_edited.plot(color='b')
        plt.show()

        df.flow2.plot(color='r')
        df.flow2_edited.plot(color='b')
        plt.show()

        df.flow3.plot(color='r')
        df.flow3_edited.plot(color='b')
        plt.show()

        df.flow4.plot(color='r')
        df.flow4_edited.plot(color='b')
        plt.show()

        df.flow5.plot(color='r')
        df.flow5_edited.plot(color='b')
        plt.show()
```





Powyższe wykresy pokazują flow w czasie deszczu dla kilku kanałów dla których z topologii można by domniemywać korelację. Niebieski kolor oznacza dane oryginalne, czerwony to miejsca które były “ręcznie” poprawiane przez specjalistów którzy uznali je za anomalie.

Wnioskiem jaki się nasuwa jest fakt że nie wiadomo czym kierował się specjalista w czasie edycji danych - widzimy ekstremalne wartości które zostały ręcznie wprowadzone ale tylko w jednym kanale albo usunięte ale też niekonsekwentnie we wszystkich kanałach. Wnioskujemy z tego że z samej topologii nie można prognozować spodziewanej reakcji na deszcz w kanałach należących do jednej topologii. Deszcz pomierzony w jednym miejscu może mieć różny obszar na którym jest obserwowany i różnie skorelowany z każdym z kanałów.
Jako wniosek z powyższego oraz podobnych wyników eksploracji danych należało odrzucić hipotezę mówiącą o automatycznym odkrywaniu topologii sieci na podstawie danych oraz tą która mówiła że można wykorzystać topologię w wykrywaniu anomalii. Pomimo tego że hipoteza ta okazała się fałszywa, możliwe było zbudowanie modelu wykrywania anomalii w oparciu o model wzorca dziennego oraz metody statystyczne.

Kamień milowy etapu 1 został osiągnięty i w etapie 3 w warunkach zbliżonych do rzeczywistych zostaną przetestowane najepsze modele predykcyjne zbudowane w etapie 1, czyli ciągła predykcja przepływu oraz wykrywanie anomalii.

Pieczęć firmowa Beneficjenta

Podpis i pieczęć osoby upoważnionej do reprezentowania Beneficjenta

Data: