Time-Series Features for Predictive Policing

Julio Borges, Daniel Ziehr, Michael Beigl
TECO, Institute of Telematics
Karlsruhe Institute of
Technology (KIT)
Karlsruhe, Germany
{lastname}@teco.edu

N. Cacho, A. Martins, A. Araujo, L.Bezerra
Federal University of
Rio Grande do Norte (UFRN)
Natal, Brazil
neliocacho@dimap.ufrn.br,
allan@dee.ufrn.br,leobezerra@imd.ufrn.br
adelsondias@ppgsc.ufrn.br

Simon Geisler
Hochschule
Albstadt-Sigmaringen
Albstadt, Germany
geislesi@hs-albsig.de

***Abstract*—Forecasting when and where crimes are more likely to occur based on years of historical record analysis is becoming a task which is increasingly helping cities' safety departments with capacity planning, goal setting, and anomaly detection. Crime is a geographically concentrated phenomena and varies in intensity and category over time. Despite its importance, there are serious challenges associated with producing reliable forecasts such as sub-regions with sparse crime incident information. In this work, we address these challenges proposing a crime prediction model which leverages features extracted from time series patterns of criminal records based on spatial dependencies. Our results benchmarked against the state of the art and evaluated on two real world datasets, one from San Francisco, US, and another from Natal, Brazil, show how crime forecasting can be enhanced by leveraging Spatio-Temporal dependencies improving our understanding of such models.**

I. Introduction

The concept of Smart City as a means to enhance the life quality of citizen has been gaining increasing importance not only for researchers but also in the agendas of policy makers. Smart cities are being piloted in the biggest cities of developed countries, such as San Francisco in the US [1], to smaller cities in developing countries such as in Natal, Brazil [2]. Smart Cities strategies are aimed at improving citizens' quality of life, coping existing challenges with increasing technological possibilities. The concept is linked to various applications from increasing citizen engagement in public policies [3], monitoring and reducing pollution [4] to enhancing public safety and city services [5] such as with "Predictive Policing", supporting police e.g. with patrol planning to make better use of limited resources . This paper focus on the latter.

Predictive Policing is "*the application of analytical techniques, particularly quantitative techniques, to identify promising targets for police intervention and prevent or solve crime*" [6]. A particular important task of Predictive Policing is forecasting criminal activities pro-actively (Crime Forecasting), i.e., predicting when and where crimes are mostly likely to occur based on analysis of historical data. Crime Forecasting however, is just that: predictions. Actual decreases in crime require taking action based on such predictions. Predictive Policing is part of an end-to-end process involving data analysts, developers and law enforcement agencies. For instance, the IEEE Smart City Initiative of Natal [2], a city of Northeastern Brazil, aims to transform Natal into a smart city through the development of systems and applications to bolster the use of IT as means of contributing to improve the life quality of its citizens. Regarding public safety, the initiative has developed the ROTA platform, a smart city platform aimed to improve public safety by integrating several information systems from different law enforcement agencies [5]. The platform delivers predictions based on historical data and actual trends, which can be further used for patrol planning supporting police operations. Designing intervention programs, combined with solid predictive analytics, can go a long way toward ensuring that predicted crime risks do not become real crimes, thus avoiding them - and Natal is just a living example of that. Predictive Policing thus depicts the shift from *reactive* policing, based on "good sense and experience" to *pro-active* policing, based on data analysis and automatized crime trend detection.

Based on the concept of Predictive Policing, this paper proposes a crime forecasting approach based on machine learning evaluated on years of historical records from two major smart cities: San Francisco (US) and Natal (Brazil). In section III, this paper proposes a pipeline of preprocessing steps in order to prepare the data for usage by machine learning regression models focusing on the application of crime forecasting such as the spatial and temporal discretization. In section IV, we propose to engineer features extracted from time series signals in order to extract patterns from crime records that can be used for time series pattern recognition, which we in turn leverage in our crime forecasting approach in section V. In the evaluation section VI, we benchmark our proposed approach against a related competitor and discuss advantages and disadvantages of our proposed method. We additionally evaluate and discuss the impact of different parametrization factors (e.g. the spatial and the temporal resolution) of our proposed approach on the crime forecast performance. Section VII then concludes this paper with section VIII focusing on future and on-going work.

978-1-5386-5959-5/18/$31.00

II. Related Work

In our focus are work related to analyzing crimes that occurred at a certain point in time at a certain place and estimate the probability of criminal activity in the future. One intuitive approach to reduce the complexity is to break down the eligible area into smaller sub-regions and treat them as atomic units. That approach takes away the spatial component and time series analysis can be performed directly on the data of the different sub-regions [7], [8], [9]. Some related work such as the one from `Brown and Oxford` [8] leverages time series with socio-economic factors – such as unemployment rates, alcohol sales and the percentage of teenagers and young adults that are living in that neighborhood – into consideration.

A related work being evaluated in this paper is the one from `Malik et al.` [7]. It is concerned with developing a framework that provides decision makers with a proactive and predictive environment based on visual analysis to assist an effective resource allocation process. They propose dividing the urban space into sub-divisons where they perform crime forecasts and underline the importance of tuning the geospatial resolution. They then apply a seasonal-trend decomposition technique based on locally weighted regression (STL) [10] to decompose the time series in their various components [7]. `Adelson et al.` [5] focus on leveraging crime forecasting based on regression techniques for patrol planning optimization, i.e., helping police patrol supervisors to elaborate a patrol planning on deciding when and which area police vehicle must patrol.

A related approach to Crime Forecasting is the Crime Hotspot Detection, the detection of sub-regions that are likely to present high criminal activities. `Borges et al.` [11] e.g. approaches a similar problem with classification rather than with regression techniques. Likewise, they are dividing the urban space into sub-regions and then they are extracting features out of the time series and the urban space (urban features) to classify whether the sub-regions are criminal hotspots.

III. Data Preparation

We employ a series of steps necessary to preprocess and to spatially discretize the data prior to being consumed by our crime forecasting model. In this section, we discuss the problem of dividing the continuous geo-space into smaller sub-regions, where crime forecasting is subsequently applied to. We subsequently describe the steps necessary to transform the criminal records into meaningful temporal data (time series) for forecasting.

A. Spatial Discretization

One real world requirement for the deployment of a crime prediction model is forecasting crimes over space and time [9]. This is a non-trivial task since the discretization of continuously distributed crimes over time and space has tremendous effects on the results and capabilities of the model. For the time dimension this is done by calculating a time series with a fixed, discrete frequency (e.g. daily, weekly, monthly) and an aggregation function (e.g. the sum of incidents). This is straightforward and well adopted in the literature.

To split continuous urban space into sub-divisions, one approach is to divide the relevant region into rectangles (grids) of the same size. This is the most adopted method in the literature [7], [8], [12]. A rectangular equisized grid is easy to calculate, to understand and draws the attention towards the application that uses the data (e.g. geo-spatial analysis or prediction). However, spatial data is most of the times not uniformly geographically distributed. Crimes for example are usually highly concentrated in the inner city and certain neighborhoods. Equisized grid cells do not account for that what can result in inaccuracies and wrong impressions of the data distribution.

Another approach is to make use man-made spatial discretization like police districts or census blocks where they exist. Police districts for example tend to get smaller in the inner city and grow in the outer regions. Nevertheless, information about police districts is not always easy to be obtained and tend not be *data-driven*, i.e., shaped in form and size by the distribution of incidents over space.

kGrid Algorithm for Spatial Discretization: we propose making use of the inherent data distribution over space and cluster regions based on their density of incidents by leveraging a soft variation of the k-Means clustering algorithm. The idea is to spatially cluster the incidents and leverage the resulting cluster's convex hulls as the grid definition, where the parameter k reflects the desired number (resolution) of grids. While for many clustering applications the parameter k is bothersome and not intuitive in this case it is a useful and easy to understand variable to control the resolution of the grid.

Figure 1 show that the clustered grid reflects the topology and crime distribution in a very useful manner. The parameter could be for example set to the number of available patrol vehicles or police units available per shift. Since it is a data oriented approach it is also possible to use different grids for different task forces depending on the category of crime. Usually, crime forecasting models are deployed for issuing forecasts for each grid individually, learning patterns and peculiarities of crime incidents in dependence of the discretization of the space.

B. Time Series Decomposition

We construct a raw time series signal out of our criminal records by taking the number of incidents (for a sub-region) over a time interval (e.g., by day, week, month). We then apply a Seasonal Trend Decomposition by Loess (STL) [10] for deconstructing the raw time series into several components, each representing one of the underlying categories of patterns, namely the derivative time series $Trend(T)$ and $Season(S)$ (plus the $Remainder(R)$). The trend component at time reflects the long-term progression of the series when there is a persistent increasing or decreasing direction in the data. The seasonal component reflects seasonality when a time series is

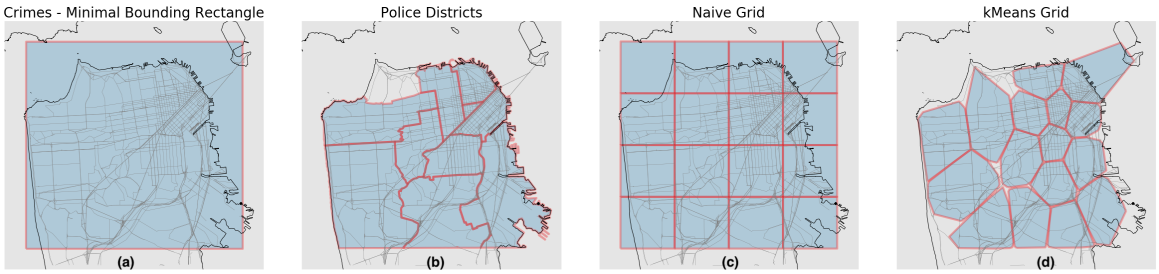

Fig. 1: Different approaches for spatial discretization of San Francisco's crime incidents. From the left to the right: **(a)** Minimal Bounding Rectangle including every incident of the dataset; **(b)** the ten Police Districts of San Francisco; **(c)** a naive rectangular grid with 16 cells; **(d)** the result of the kGrid (kMeans Grid) algorithm with 16 cells.

influenced by seasonal factors. The remainder expresses the left-over noise within the data. Hence, a time series (Y) at time t using an additive model as the STL suggests can be thought of as:

$$Y_t = T_t + S_t + R_t \tag{1}$$

The result is the split of the input time series of crime records into the two derivatives Trend and Season, from which Time Series Features (IV) are also extracted. We refer to [10] for details about the STL functioning.

IV. Time Series Features (TSF)

One contribution of this work is to examine the importance of different time series features for the prediction of crimes. In this section it will be explained which features are extracted as well as the methodology of obtaining them. The underlying time series of this work is the number of crime incidents for a given time interval (time step: e.g. weekly) and geographic region (cf. III-B).

Definition 4.1 (Time Series Feature (TSF)): A Time Series Feature (TSF) is a function that takes the past `n_feat` observations of a time series and maps it on a single, numeric value.

Our feature extraction mainly utilizes the python package `tsfresh` [13] that automatizes such time series feature extraction. The list of features that are extracted can be complemented by user defined functions. For the following definitions the underlying time series is denoted as y_t at time t. The span that is used to calculate the features is denoted as `n_feat`.

TSF 1 (Absolute Energy): The Absolute Energy of the time series y is defined as the sum over the squared values:

$$E = \sum_{t=1,\ldots,\texttt{n\_feat}} y_t^2 \tag{2}$$

TSF 2 (Mean Change): The Mean Change returns the average differences between subsequent time series values and is defined as

$$\frac{1}{\texttt{n\_feat}} \sum_{t=1,\ldots,\texttt{n\_feat}-1} y_{t+1} - y_t \tag{3}$$

TSF 3 (Mean Absolute Change): The Mean Absolute Change returns the average absolute difference between subsequent time series values y_t and y_{t+1} and is defined as

$$\frac{1}{\texttt{n\_feat}} \sum_{t=1,\ldots,\texttt{n\_feat}-1} |y_{t+1} - y_t| \tag{4}$$

TSF 4 (Augmented Dickey-Fuller): The Augmented Dickey-Fuller is a hypothesis test which checks whether a unit root is present in a time series sample. This feature calculator returns the value of the respective test statistic.

TSF 5 (Index Mass Quantile): The Index Mass Quantile is defined as the relative index i where $q\%$ of the mass of the time series y lay left of i. For example for $q = 50\%$ this feature will return the mass center of the time series. The mass of the time series is defined as the cumulative sum. The values that is used are $y_{t-\texttt{n\_feat}}$ to y_t

TSF 6 (Cross Power Spectral Density): This feature calculator estimates the cross power spectral density of the time series x at different frequencies. To do so, first the time series is shifted from the time domain to the frequency domain. The feature calculators returns the power spectrum of the different frequencies.

TSF 7 (Standard Deviation): Returns the standard deviation of time series x.

TSF 8 (Time Reversal Asymmetry Statistic): This function calculates the value of

$$\frac{1}{\texttt{n\_feat} - 2lag} \sum_{i=0}^{\texttt{n\_feat}-2lag} y_{i+2\cdot lag}^2 \cdot y_{i+lag} - y_{i+lag} \cdot y_i^2 \tag{5}$$

which is $\mathbb{E}[L^2(Y)^2 \cdot L(Y) - L(Y) \cdot Y^2]$ where $\mathbb{E}$ is the mean and L is the lag operator. It was proposed in [14] as a promising feature to extract from time series.

TSF 9 (Continuous Wavelet Transformation Coefficients): Calculates a Continuous wavelet transform for the Ricker wavelet, also known as the "Mexican hat wavelet" which is defined by

$$\frac{2}{\sqrt{3a}\pi^{\frac{1}{4}}}\left(1 - \frac{y^2}{a^2}\right)\exp\left(-\frac{y^2}{2a^2}\right) \tag{6}$$

where a is the width parameter of the wavelet function.

In total, we leverage over 100 Time Series Features, listing above just the most important ones with impact on our model.

We further refer to [13] for a detailed overview of all leveraged features.

Another basic category of time features are the lag functions.

Definition 4.2 (Lag (TSF-LAG(n)): Simple lag function that shifts the time series by n steps so that

$$\text{TSF-LAG}(n) = y_{t-n} \tag{7}$$

Lagged features represent the raw time signals at a certain time step t.

V. Crime Forecasting Model

We pose the crime forecasting problem as a multivariate regression problem, where the independent variables are derived from our proposed Time Series Features (TSFs).

In this section, we tap the task of selecting the right machine learning regression model for our task at hand (V-A). We then discuss how to minimize the complexity of the model by taking only important features and eliminating correlated ones in sec. V-B. We finally describe our whole crime forecast process in sec. V-C.

A. Regression Model Selection

We leverage 3 well adopted regression models for this task, namely: Support Vector Regression (SVR), Multi Layer Perceptron Regression (MLPR) and Random Forest Regression (RFR). Every model has a different mathematical base and brings certain advantages and disadvantages with it. Especially the possibility to access the feature importance is of high interest for this work.

In order to enable a fair comparison between the 3 regression models, we benchmark them by fitting each model with training data consisting of three years (156 weeks) of crime records and testing it on a subsequent year (52 weeks) for a pre-selected region of San Franscisco. To make the testing results more robust, cross validation for time series is applied. While classical cross validation assumes the independence and identically distribution of samples between each other, time series cross validation has to account for the autocorrelation of the time series data. The most important difference is that successive training sets are supersets of those that came before them. Table below shows the performance of the regression models for 3 for the Mean Square Error (MSE). We set the spatial resolution fixed and performed a Grid Search on the parameters of all 3 algorithms to minimize the error function (the MSE).

Measure / Model	SVR	MLPR	RFR
MSE	961.326	1718.571	897.149

TABLE I: Model Selection: performance of 3 regression algorithms for the task of crime forecast

The Random Forest Regressor (RFR) outperformed both the SVR and MLPR in our tests scenario and will be used as the main regression algorithm underlying our crime forecasting method.

B. Recursive Feature Elimination (RFE)

After extracting the multiplicity of TSFs (see def. 4.1), we prune (eliminate) features with lower importance which have no impact on the regression model. The importance of a feature is computed as the (normalized) total reduction of the criterion brought by that feature, also known as the Gini importance [15]. The features with the lowest importance are recursively pruned until the desired number of features is reached. This external parameter will later be optimized to find the best number of features that should remain. Another reason for the deployment of RFE is its robustness against highly correlated features. In their work `Gregorutti et al.` [16] examined the problem of feature selection with the Random Forest algorithm in high-dimensional regression settings. Especially for highly correlated features - such as temporal ones (TSFs) - they recommend the use of RFE for feature selection [16].

C. Crime Forecasting Process

After describing the individual parts of the Crime Forecasting Model, they will now be put together and considered in a broader context. The Crime Prediction Process describes the steps necessary to obtain meaningful crime predictions (see figure 2). We now list the steps involved in our geospatial prediction methodology:

1) **Dividing geospace into sub-regions:** discretize the geospace in grids where to forecast crimes individually
2) **Generating the time series signal:** apply Season Trend Decomposition (STL) and calculate Time Series Features (TSFs) based on the crime incident records from each sub-region.
3) **Regressor Fitting:** train (fit) a Random Forest Regressor for each grid based on a subset of TSFs using recursive feature elimination (RFE).
4) **Forecasting:** The time series features generated for each spatial unit (grid) is then fed through the RFR process where a forecast is generated for the next N time steps (weekly). This process is repeated for all region subdivisions (grids) and prediction maps are finally obtained for the next N time steps.

The parameters of the model are summarized in table II. A particular parameter of impact is the spatial resolution, i.e., the number of grids (sub-regions) to be generated (parameter k of $kGrid$). A high value for k may result in a grid resolution that is too fine, generating a zero count vs. time step signal that has no predictive statistical value. On the other side, a grid resolution that is too coarse (small k) may introduce variance and noise in the input signal, thereby over-generalizing the data. As denoted by `Malik et al.` [7] an average input size of 10 samples per time step provides enough samples for extracting meaningful patterns from time series as rule of thumb. W.r.t $category$ and $frequency$ (cf. Table II), this paper focuses on violent felonies with a weekly forecast prediction

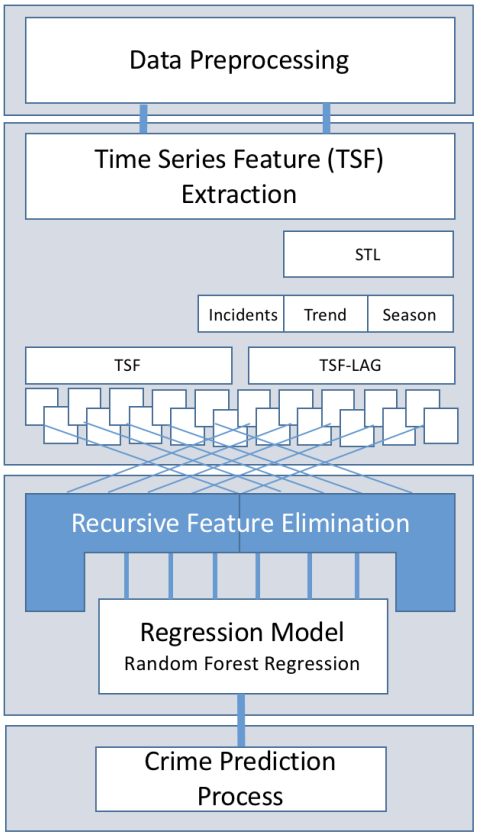

Fig. 2: Summary of proposed Crime Forecasting Model

horizon and weekly time steps. Lower time steps (e.g. daily) turned out not to be practical in our experiments, generating many sparse grids with less than 10 samples per time step even for lower k values.

We vary the amount of data we use to train the regressor model (RFR) between 156 - 280 weeks and choose to keep between 3 - 60 time series features depending on other parameters setting. We then optimize the overall parameter selection through Grid Search [17] finding the best set of parameters for our task at hand.

We must note however, that our method assumes spatial independency of observations, i.e., for each sub-region we train and apply a model only trained with data from that sub-region without considering patterns from neighboured regions. It is however expected the activity on one criminal area can affect another. We plan to research and address such effects in future work (cf. sec. VIII).

VI. Evaluation

We evaluate our approach on two datasets. One from San Francisco (US) with 12 years (2003-2015) of crime records and on one from Natal (Brazil) with 10 years (2006-2016) of crime records. We focus on violent felonies in both datasets instead of all crime categories in general. We evaluate and discuss the impact of several parameters of our proposed crime forecasting based on consistent evaluation metrics.

A. Evaluation Metrics

The main evaluation metrics used in this work is the Mean Squared Error (MSE) defined as average squared deviation between observation and forecast, given by:

$$MSE(y, \hat{y}) = \frac{1}{n}\sum_{t=1}^{n}(y_t - \hat{y}_t)^2 \tag{8}$$

And the Mean Absolute Percentage Error (MAPE), defined as average absolute deviation between observation and forecast normalized by the observation, given by:

$$MAPE(y, \hat{y}) = \frac{100}{n}\sum_{t=1}^{n}\left|\frac{y_t - \hat{y}_t}{y_t}\right| \tag{9}$$

Whereby y stands for the observed time series of length n and $\hat{y}$ stands for the forecasted values.

MAPE is well suited for our purposes due to the fact that it does not consider absolute but rather relative values, being suitable for means of comparison without revealing absolute information on the underlying data.

B. Forecast Performance

In order to compare the performance of our approach to the state-of-the-art, we developed and implemented the crime forecasting approach proposed by `Malik et al.` [7] to serve as a baseline. We parametrized the baseline to minimize forecasting error functions described previously through Grid Search. We train both models on 4 years of crime incident records and test it on 2 years of crime records from both cities: San Francisco and Natal. We focus on weekly predictions and discretize the geospace into 22 grids, since these parameters delivered the best performance as later discussed.

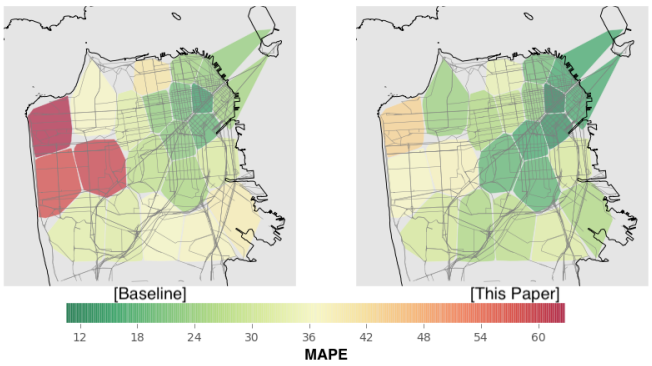

Fig. 3: Prediction results for the city of San Francisco. Especially the sparse grids in the western part of the city have shown strong improvements.

The results in figure 3 show that our model outperforms the baseline in every single cell. This is true throughout the whole city. It can also be seen that the most perceptible improvement happens in the western part of the city, where less criminal activity takes place, leading to sparse time series signals.

Name	Description	Value
`category`	Specifies the Crime Group Category to be forecasted.	`Violent Felonies, etc. -`
`frequency`	Temporal frequency of the time series.	`weekly`
`k`	Determines number of cells for the kGrid algorithm	
`n_feat`	Span used for TSF Extraction. Also indirectly determines the number of TSFs that are generated and passed to the RFE algorithm	
`n_feat_select`	Number of features that are selected by the RFE algorithm.	`3,...,60`
`n_train`	Number of periods used to fit the model.	`156,...,280`
`n_test`	Number of periods used to test the model.	`56,104`

TABLE II: Parameters of the Crime Prediction Model with a short description. Value column shows exemplary parameter choices or the range we used throughout evaluation (in weeks if numeric).

That means our model especially outperforms the baseline for sparse data.

The next generalization step is to test our model on a second data set - the one from Natal, Brazil. As there are few data available for Natal compared to San Francisco, we performed a second Grid Search for optimize our model according the best parameters minimizing error functions. The results are shown in figure 4 and support in general the findings from San Francisco. Although for Natal there are more regions which don't improve or even get worse results for our model when compared to the baseline.

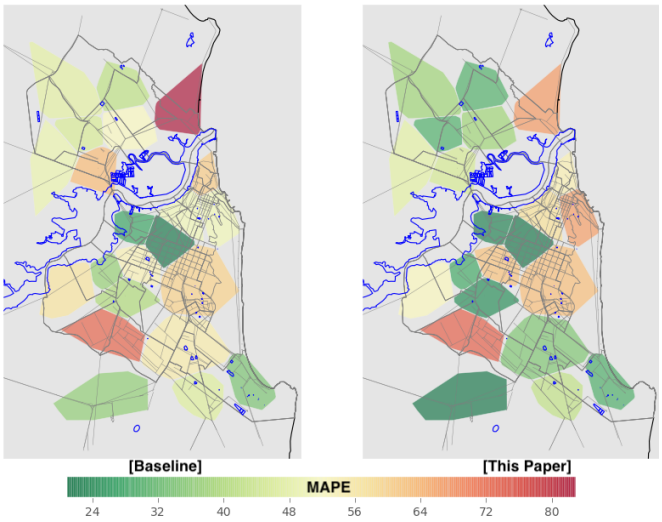

Fig. 4: Prediction results for the city of Natal on a grid of 22 cells. Improvement compared to the baseline, especially in the north of the city.

In order to prove that our results are statistically better (significant) we perform a Kolmogorov-Smirnov-Test (KS-Test) comparing our results for every grid against the baseline. With the KS-Test it is possible to compare two samples and determine if they are from the same distribution. This would be one way to statistically proof that our model performs better. The null hypothesize is that the underlying distributions are identical. With a p-value of 0.035 we can reject the null hypothesizes for violent felonies in San Francisco for $\alpha = 0.05$. That means our model is statistically significant better than the baseline for this crime category. We would like to note that for some crime categories not further investigated in this work the null hypothesis could not be rejected. That, in turn, does not mean that the underlying distributions are the same. But it is an important limitation of this evaluation's results and show necessity of future work on the predicability of different crime categories and the generalisation of such forecasting models across different crime categories.

C. Varying the Spatial Resolution

As discussed previously, the choice of the spatial resolution has a great impact on the Crime Forecasting task. It determines the geographical resolution of the prediction and also influence the quality of the prediction within each grid. Those quality aspects are opposing each other. In this section, we will analyze the quality and behavior of our model for varying the parameter `k` of the kGrid algorithm. A qualitative approach is to plot the different grid sizes and compare their results visually. This gives a very good impression on how the grid evolves with growing `k` and how our model behaves compared to the baseline.

Figure 5 shows for different `k` how the resulting forecasting results evolves for the city of San Francisco. The higher the `k` becomes, the more differentiated the results are. While for small `k` the differences of the results are smaller, the differences for higher `k` become observable. This shows how our algorithm perform extremely well when compared to the baseline on higher spatial resolutions.

D. Time Span for Feature Calculation

The Time Series Features (see def. 4.1) are defined as functions that maps the past part of a time series on a numeric value. The length of this span is determined by the model's parameter `n_feat` (cf. tab. II). The higher this parameter is the more information about the crime history can be incorporated. On the other hand short term changes might be missed if the parameter is too large. We evaluated the time span on a grid of 22 cells and for the violent felonies of San Francisco, by varying the time span from 1 to 4 years. Our results suggest that shorter spans are generally better suited than longer spans and we keep one year time span as default for our parameter `n_feat`.

E. Number of Features

Another interesting question is how many TSFs should be selected in the RFE step. The RFE algorithm (see sec. V-B) eliminates features until the the external number

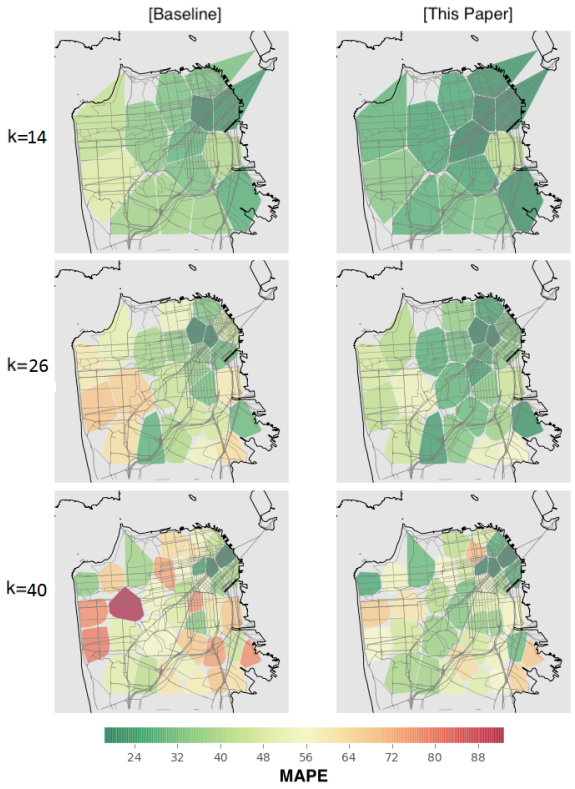

Fig. 5: Prediction results for violent felonies in San Francisco with different grid sizes. It can be seen how our proposed approach outperforms the baseline in all tested cases, especially for higher spatial resolutions.

`n_feat_select` (cf. table II) is left. Intuition suggests that the more features we are taking into consideration the more information is passed to the model and thus the quality improves. Contrarily to this intuition it can be shown that taking too many features into consideration can impair the results. The high dimensionality adds complexity and noise to the model. In high dimensional space, patterns and characteristics of data can blur away. To have a deeper look into the behavior of our model for different numbers of TSFs, a series of prediction runs with varying number of features to select is conducted.

In our experiment settings, our model outperformed the baseline independently of the number of features that have been selected on the San Francisco dataset, by varying the number of features from 10 to 60. However, our empirical observation suggests that sparse grids tend to have better results with less features (around 10).

F. Feature Importance

To evaluate the importance of individual TSF over a whole grid we are proposing the Feature Cell Count Ratio (FCCR):

Definition 6.1 (Feature Cell Count Ratio (FCCR)): The Feature Cell Count Ratio (FCCR) is expressing the importance of an individual Time Series Feature (TSF) over a grid. It is calculated as the number of occurrences in different cells over the grid. That means every time TSF f is not eliminated by the RFE algorithm (see sec. V-B) the feature cell count $|f|$ of f is increased by one. After that the count is normalized by the number of cells in the corresponding grid. If k denotes as the number of cells in the grid, the FCCR is defined as:

$$FCCR(f) = \frac{|f|}{k} \tag{10}$$

That means $FCCR(f) = 1$ if f is present in every cell of the grid and decreases until $FCCR(f_i) = 0$ if f is not present in any cell.

We then evaluate the feature importance for both cities for violent felonies with parameter $k = 22$. The most important (respectively most selected) feature is the seasonal component lagged by one year [TSF-LAG(53)]. Another interesting peculiarity is that both configurations have the Absolute Energy calculated on the Time Series Signal and on the Trend Signal. That points out that the movement over the past period is well suited to predict future outcomes. To summarize this section it can be said that the intersection of TSF-sets for the different cities is surprisingly large. These few TSF (compared to the number of TSF that are extracted) are dominating the top selected features. This is a valuable finding for future work on this topic.

San Francisco

Category	Basis	Description	FCCR
TSF-LAG(53)	Seasonal	Shifted by 53 weeks (1 year)	1.000
TSF-LAG(1)	Seasonal	Shifted by 1 week	0.636
TSF	Crimes	Absolute Energy	0.318
TSF	Seasonal	Cross Power Spectral Density	0.273
TSF	Trend	Absolute Energy	0.273
TSF-LAG(1)	Crimes	Shifted by 1 week	0.273
TSF-LAG(53)	Crimes	Shifted by 53 weeks (1 year)	0.273
TSF-LAG(14)	Seasonal	Shifted by 14 weeks	0.136
TSF	Seasonal	Cross Power Spectral Density	0.136
TSF	Seasonal	TRAS[1], Lag 1	0.136

Natal

Category	Basis	Description	FCCR
TSF-LAG(1)	Seasonal	Shifted by 1 week	0.882
TSF	Seasonal	Cross Power Spectral Density	0.765
TSF	Seasonal	Mean Change	0.647
TSF	Trend	TRAS, Lag 1	0.647
TSF	Trend	Absolute Energy	0.529
TSF	Trend	Mean Change	0.471
TSF-LAG(53)	Trend	Shifted by 53 weeks (1 year)	0.412
TSF	Crimes	Absolute Energy	0.412
TSF	Trend	TRAS, Lag 3	0.235
TSF	Seasonal	Cross Power Spectral Density	0.176

TABLE III: The table shows the FCCR (cf. def. 6.1) of the most important TSFs.

VII. Conclusion

In this paper, we proposed a machine learning based framework for the task of crime forecasting. A particular contribution of this paper lies of the demonstration of features which can be extracted from time series signals leveraged by machine learning models for this task. We have shown

that some time series features such as the *Absolute Energy* or the *Cross Power Spectral Density* are particularly relevant, independent of the dataset (City) being explored.

We additionally delivered insights on how to preprocess such datasets for optimal consumption by machine learning models based on time series decomposition, feature importance and elimination and on evaluating the parametrization of the spatial and temporal discretization on the forecasting performance.

We evaluated our proposed approach on datasets containing over 10 years of criminal records from 2 cities: San Francisco, US and Natal, Brazil – benchmarking our proposed approach to the state of the art. Our results show that our approach not only outperforms the evaluated competitor for this task but is also less sensitive to spatio-temporal resolution performing particularly well in sparse areas – i.e., regions of the space with sparse incident information over time.

We believe to make a significant contribution to Predictive Policing initiatives, such as helping law enforcement agencies to make better usage of limited resources such as with better patrol planning based on the most likely predicted criminal areas with our approach.

VIII. Future Work

One limitation of crime forecasting approaches in related work (e.g. [7]) and the one proposed by this paper regards the assumption of spatial independence of observations: for each geographical sub-region (grid) a crime forecasting model is independently trained and applied to that area. It is however expected the activity on one criminal area can affect another. This is a common phenomenon in geographic data and it can be interpreted as direct demonstration of Tobler's First Law of Geography: "*everything is related to everything else, but near things are more related than distant things*" [18]. It is still a open research question how general such models can be with respect to geographic regions, i.e., trained with data from one region and applied to another. And how to leverage information of another regions to account for spatial correlation. It is also not known how such models generalize to different crime categories. Additionally, regression techniques require the independence of observations; however, some features might show spatial auto-correlation (i.e., spatial dependency). In practice, spatial auto-correlation is the tendency of nearby observations to be correlated to one another. And this can be tested with spatial correlation tests. We plan to address such questions in near future work.

[1]Time Reversal Asymmetry Statistic

References

[1] J. H. Lee, M. G. Hancock, and M.-C. Hu, "Towards an effective framework for building smart cities: Lessons from seoul and san francisco," *Technological Forecasting and Social Change*, vol. 89, pp. 80–99, 2014.

[2] N. Cacho, F. Lopes, E. Cavalcante, and I. Santos, "A smart city initiative: The case of natal," in *2016 IEEE International Smart Cities Conference (ISC2)*, Sept 2016, pp. 1–7.

[3] J. Borges, M. Budde, O. Peters, T. Riedel, and M. Beigl, "Towards two-tier citizen sensing," in *Smart Cities Conference (ISC2), 2016 IEEE International*. IEEE, 2016, pp. 1–4.

[4] M. Budde, R. El Masri, T. Riedel, and M. Beigl, "Enabling low-cost particulate matter measurement for participatory sensing scenarios," in *Proceedings of the 12th International Conference on Mobile and Ubiquitous Multimedia*, ser. MUM '13. New York, NY, USA: ACM, 2013, pp. 19:1–19:10.

[5] A. A. Junior, N. Cacho, A. C. Thome, A. Medeiros, and J. Borges, "A predictive policing application to support patrol planning in smart cities," in *2017 International Smart Cities Conference (ISC2)*, Sept 2017.

[6] W. L. Perry, *Predictive policing: The role of crime forecasting in law enforcement operations*. Rand Corporation, 2013.

[7] A. Malik, R. Maciejewski, S. Towers, S. McCullough, and D. S. Ebert, "Proactive spatiotemporal resource allocation and predictive visual analytics for community policing and law enforcement," *Visualization and Computer Graphics, IEEE Transactions on*, vol. 20, no. 12, pp. 1863–1872, 2014.

[8] D. E. Brown and R. B. Oxford, "Data mining time series with applications to crime analysis," in *Systems, Man, and Cybernetics, 2001 IEEE International Conference on*, vol. 3. IEEE, 2001, pp. 1453–1458.

[9] C.-H. Yu, M. W. Ward, M. Morabito, and W. Ding, "Crime forecasting using data mining techniques," in *Data Mining Workshops (ICDMW), 2011 IEEE 11th International Conference on*. IEEE, 2011, pp. 779–786.

[10] R. B. Cleveland, W. S. Cleveland, J. E. McRae, and I. Terpenning, "Stl: A seasonal-trend decomposition procedure based on loess," *Journal of Official Statistics*, vol. 6, no. 1, pp. 3–73, 1990.

[11] J. Borges, D. Ziehr, M. Beigl, N. Cacho, M. Martins, S. Sudrich, S. Abt, P. Frey, T. Knapp, M. Etter, and J. Popp, "Feature engineering for crime hotspot detection," *IEEE International Conference on Smart City Innovations (IEEE SCI 2017)*, 2017.

[12] R. Maciejewski, R. Hafen, S. Rudolph, S. G. Larew, M. A. Mitchell, W. S. Cleveland, and D. S. Ebert, "Forecasting hotspotsa predictive analytics approach," *IEEE Transactions on Visualization and Computer Graphics*, vol. 17, no. 4, pp. 440–453, 2011.

[13] M. Christ, A. W. Kempa-Liehr, and M. Feindt, "Distributed and parallel time series feature extraction for industrial big data applications," *arXiv preprint arXiv:1610.07717*, 2016.

[14] B. D. Fulcher and N. S. Jones, "Highly comparative feature-based time-series classification," *IEEE Transactions on Knowledge and Data Engineering*, vol. 26, no. 12, pp. 3026–3037, 2014.

[15] L. Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.

[16] B. Gregorutti, B. Michel, and P. Saint-Pierre, "Correlation and variable importance in random forests," *Statistics and Computing*, pp. 1–20, 2013.

[17] M. Claesen and B. De Moor, "Hyperparameter search in machine learning," *arXiv preprint arXiv:1502.02127*, 2015.

[18] W. R. Tobler, "A computer movie simulating urban growth in the detroit region," *Economic geography*, vol. 46, no. sup1, pp. 234–240, 1970.