Software Engineering Department
Ort Braude College
Course 61401: Extended Project in Software Engineering

Prediction and Forecasting Crime Rates with Gaussian Processes

In Partial Fulfillment of the Requirements for
Final Project in Software Engineering (Course 61771)
Karmiel

Elad Fishman
Gal Menashe

Supervisor: Prof. Valery Kirzner
Advisor: Prof. Zeev Volkovich

Contents

1. Introduction 3

2. Background and Related Work 4

2.1 Poisson Distribution 4

2.2 Gaussian Process 4

2.2.1 Prediction 5

2.3 Gaussian Regression 6

2.4 Covariance Function 6

2.5 Heat Map 6

2.6 Tools Used 6

3. Our Model 6

3.1 Database 8

4. Preliminary Software Engineering Documents 10

4.1 Flow Diagram 10

4.2 Requirements (Use Case) 11

4.3 Design (GUI, UML Diagrams) 11

4.3.1 GUI And User Manual 12

4.4 Testing Plan 14

5. Results and Conclusion 15

5.1 Results 15

5.1.1 Choosing the Covariance Function 15

5.1.2 The Method of Drawing the Colors Intensity 17

5.1.3 Examination of The Heatmap 18

5.2 Conclusion 18

6. References 19

1. INTRODUCTION

One of the major problems around the world is crime. One proposed definition for crime is a harmful act not only to some individual but also to a community, society or the state. Such acts are forbidden and punishable by law and have made us trying to figure out what is causing this phenomenon. A single crime case probably arises from the individual thought of a person, for that reason, it seems impossible to forecast who will commit a crime, but can the rate of crime cases in a selected location be predicted?

In the last decades the number of crime cases relative to population size has increased, as opposed to expectations. Due to the increase in the amount of crimes, the police found itself under pressure and in searching after alternatives to minimize crime cases. One idea that has come up and is in a great development momentum is predicting crimes based on analyzed previous crime cases.

In order to decrease crime rates, the police have established an extensive database that was based on large amounts of data collected from previous crimes, and cataloged it by location, time and type of crime. This database will be used in our project, which propose an algorithm that will be used to predict crime cases. The most common method of crimes prediction is to assume that "hotspots", which are found with the kernel strength estimate, will remain in the short term. Other attempts included prediction using extrapolation based on time series.

This project focuses on the geostatistical map of crimes and tries to create a framework that will be flexible and based on Gaussian processes for modeling and short-term prediction of space and time crime with three objectives: prediction and evaluation, spatial focus and temporary focus. The model designed to be flexible data-driven, it does not take into account the dynamics of crime, it takes the amount of crimes as data for the statistical problem of accurate forecasting models criminal counts in space and time. The model is quite general and can be transformed to use other spatiotemporal data.

This project will use the Gaussian process model and examine which covariance function gives the most accurate prediction results. In order to do the examination, we will study the rate of crimes in the previous years, including the last year given, and then predict the crime rate of the last given year with a number of different covariance functions, the last step is to check which of the covariance functions gives the most accurate prediction. In practice, we will study the crime rate of 2013-2017, then predict the last year that is given 2017 several times, each time with other covariance function and compare it to the actual data until we reach the covariance function that will give the most accurate prediction.

2. BACKGROUND AND RELATED WORK

2.1 Poisson Distribution

The Poisson distribution is a discrete probability distribution that expresses the probability of a given number of events occurring in a fixed interval of time or space, if these events occur with a known constant rate and independently of the time since the last event. The Poisson distribution can also be used for the number of events in other specified intervals such as distance, area or volume.

$$p(k) = \frac{\lambda^k}{k!} e^{-\lambda}$$

λ is the average number of events per interval, is also called as event rate.
k is the number of times an event occurs in an interval and it can only take positive integer values 0,1,2,..

2.2 Gaussian Process

Gaussian process is a stochastic process with a collection of random variables indexed by location and time, any finite set of these random variables has multivariate normal distribution. The distribution of a Gaussian Process is the joint distribution of all those random variables.
A Gaussian process realization is a function $f(x)$ and each run of the process will provide a different result due to its stochastic attribute.
A Gaussian process is specified by its mean function $\mu(x)$ and its covariance function $k(x, x')$, a variety of covariance functions can be chosen to be used in the Gaussian process in order to obtain different aspect of the results. [3]

$$f \sim GP(\mu(x)\, , k(x, x'))$$

A Gaussian process is a complex process. For example, an illustration of the result of a less complex run, by taking mean equal to zero, $\mu(x) = 0$ and the covariance function $k\left(x_i, x_j\right) \; = \; exp(\left\|x_i - x_j\right\|^2)$. In practice, creating a grid of points X which will accept values between 2 and -2, and calculating the covariance K, next step is drawing Y from a multivariate Gaussian distribution with mean 0 and covariance K. Three different runs of the Gaussian process will give different results, as we can see in Figure 1.

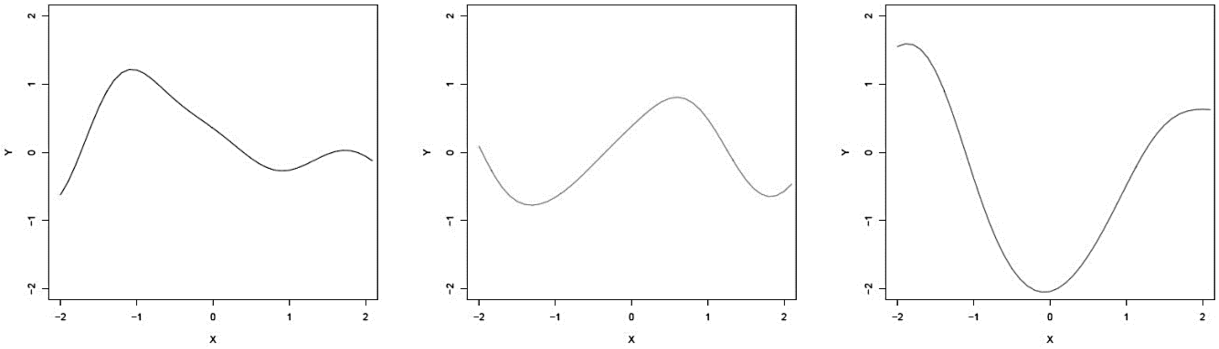


Figure 1: Gaussian process example

2.2.1 Prediction

The example above, figure 1, is a randomly drawn functions with Gaussian processes. In order to use Gaussian processes, we want to include the knowledge that the training data provides within the process. The joint distribution over both training outputs, y, and the test outputs y_* according to the prior is:

$$\begin{pmatrix} y \\ y_* \end{pmatrix} \sim N\left(\mu(x), \begin{bmatrix} K(X,X) & K(X,X_*) \\ K(X_*,X) & K(X_*,X_*) \end{bmatrix}\right)$$

Where X represent the observed inputs and X_* the unobserved.
The conditional distribution can be found according to the observed (X,Y) using the properties of multivariate Gaussian distributions.

$$Y_*|Y \sim \mathrm{N}(K(X_*,X)K(X,X)^{-1}Y, K(X_*,X_*) - K(X_*,X)K(X,X)^{-1}K(X,X_*))$$

This is simple case where the observations are noise free, for more realistic modelling situations we interesting to use the case with the noise. This affects the covariance diagonal where $i = j$.

$$Y_*|Y \sim \mathrm{N}(K(X_*,X)(K(X,X) + \sigma^2 I)^{-1}Y, K(X_*,X_*) - K(X_*,X)(K(X,X) + \sigma^2 I)^{-1}K(X,X_*))$$

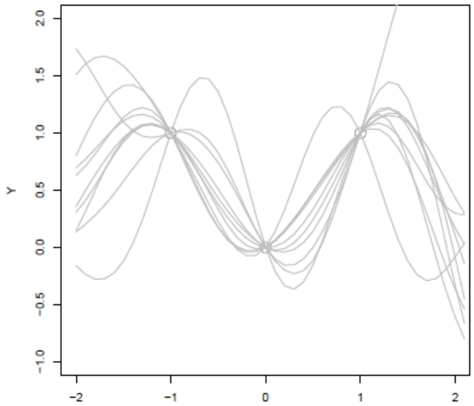


Figure 2: Draws from a GP posterior without noise variance

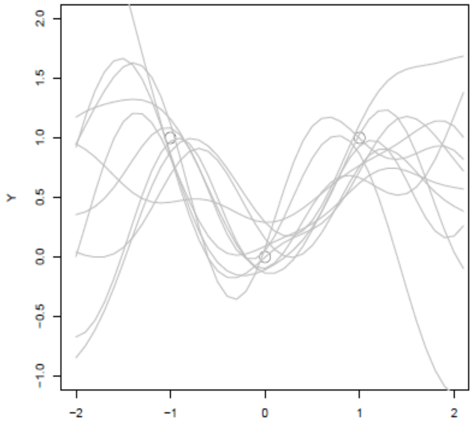


Figure 3: Draws from a GP posterior with noise variance

2.3 Gaussian Regression

The application of Gaussian processes is called Gaussian regression and it's particularly useful for problems without a known special model function. The method has features of a machine learning procedure and enables automatic modeling based on observations. A Gaussian process captures the typical behavior of the system, which can be used to derive the optimal interpolation for the problem. The result is a probability distribution of possible interpolation functions as well as the solution with the highest probability.[5]

2.4 Covariance Function

Covariance is a measure of the similarity between two or more random variables. Covariance is positive when variables tend to behave similarly, above or below average, and negative when they change in opposite directions. Covariance function describes the spatial or temporal covariance of a random variable process or field, and it is the main part using of Gaussian processes. There are many ways to construct covariance function: sum, product and convolution.

2.5 Heat Map

To represent the prediction results, we use the "Heat map" method, which helps visually present and analyze the data more efficiency than directly observing the results of the predictions. The predicted crime locations can be seen as more intense colored area on the map. [8]

2.6 Tools Used

We have used python 2.7 interpreter with the main packages ”Matplotlib” for the graphs, “TKinter” for the GUI, and “Sklearn”[1] Python’s peerless machine learning library. It provides a comprehensive set of supervised and unsupervised learning algorithms, implemented under a consistent, simple API that makes the entire modeling pipeline as frictionless as possible, a GaussianProcessRegressor is applied by specifying an appropriate covariance function. and applied by specifying an appropriate covariance function it does not allow for the specification of the mean function, always assuming it to be the zero function, highlighting the diminished role of the mean function in calculating the posterior.

3. OUR MODEL

Gaussian process provides a prior for functions with a set of values distributed by multivariate Gaussian distribution. Meaning, it will be used as a continuous data model with real-valued data. Count data 'y' is positive, and integer valued, this type of data is usually modeled with a Poisson distribution:

$$p(y|\lambda) = \frac{\lambda^y}{y!} e^{-\lambda}$$

The underlying rate parameter λ needs to be estimated with exponential function according to history of crimes, considering space and time $\lambda(s,t)$. The estimation will be based on known results of historical data analysis. The dependence on space and time cause $\lambda(s,t)$ to be positive real-valued function, so we place the log of the function with a Gaussian process prior $f(s,t)$:

[1] http://scikit-learn.org/stable/modules/gaussian_process.html

$$\lambda(s,t) = \exp\big(f(s,t)\big)$$

Following the location of the historical crimes there will be including a spatial term for the expected count at location s, called e_s:

$$y_{s,t}|\lambda(s,t) \sim Poisson(\exp\big(f(s,t)\big) \cdot e_s)$$

The part $\exp\big(f(s,t)\big)$ interpret as the relative risk and $f(s,t)$ as the log relative risk. When $f(s,t) = 0$, $\exp\big(f(s,t)\big) = 1$, so $y_{s,t}$ gets a Poisson distribution with mean equal to the expected count at location s, e_s. When $f(s,t) > 0$, $\exp\big(f(s,t)\big) > 1$ so $y_{s,t}$ gets mean that greater than e_s and when $f(s,t) < 0$ the mean reduced. Conditional on the relative risk surface f, the expected counts are independent, so the likelihood factors:

$$p(y|f) = \prod_{s,t} Poisson(y_{s,t}|\exp\big(f(s,t)\big) \cdot e_s)$$

The hidden surface f will be replaced with Gaussian process function where the mean is 0 and covariance is K:

$$f \sim GP(0,K)$$

In order to complete the formula, the last thing left to examine is which covariance function will give the most accurate prediction results according to the analysis of previous crime cases and predicting a period from the past, which is also given.
We examined several options of covariance K:

The first covariance function depends on space and time and its structure will be built from combining spatial covariance $k_s(s,s')$ and temporal covariance $k_t(t,t')$.

$$K(i,i') = k_s(s,s') + k_t(t,t')$$

The second covariance function depends on a periodic temporal term $k_p(t,t')$, to account for seasonal variations, there are many variations to represent this function and we will use the following parameterization [3]

$$k_p(t,t') = \exp\left(\frac{2\sin^2\left(\frac{(t-t')\pi}{52}\right)}{l^2}\right)$$

The third covariance function will contain the combination of the two previous ones, and thus we will try to create a covariance function that will take into account most factors and may give a more accurate result than the previous functions:

$$K\Big((s,t),(s',t')\Big) = k_s(s,s') + k_t(t,t') + k_{st}\Big((s,t),(s',t')\Big) + k_p(t,t')$$

Deriving the conditional distribution corresponding we arrive at the key predictive equations for Gaussian process regression

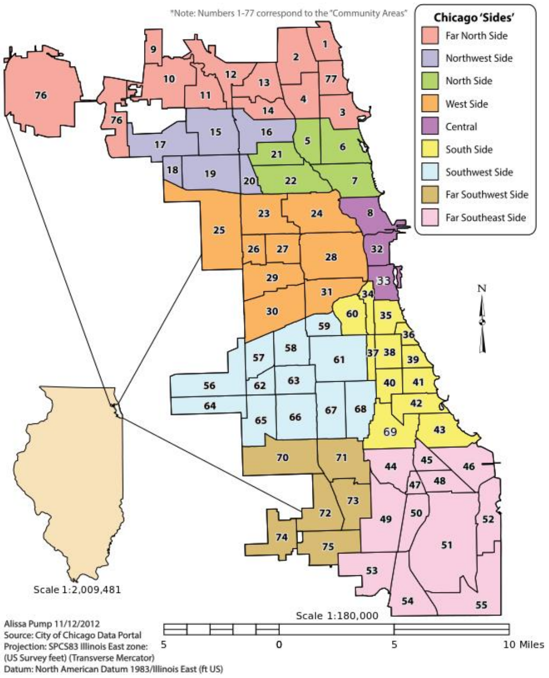
$$p(y_*|y) = N(0,K_*)$$

$$K_* = K(X_*,X_*) - K(X_*,X)(K(X,X) + \sigma^2 I)^{-1}K(X,X_*)$$

3.1 Database

We chose to analyze Chicago crime rate on the algorithm. First, we downloaded the database (Table 1: Database from the website CityOfChicago) which the project is based on. This database is public and maintained by the Chicago authorities[2], it's available through the data portal. In the algorithm the database will contain samples from 1/1/2013 to 1/6/2018 and will be aggregated by type of crime, neighborhood (Chicago is divided into 77 community areas Figure 4: Chicago Community Areas Map and Table 2: Database table aggregated by neighborhood and week) and the week of the year. Since the number of all the documented crimes is enormous and due to resource limitations for computation, the type of crime that we investigated in the project is grouped according to five common crimes (Figure 5) - which makes results modelling easier.

Crimes are being committed all over the world every day and there is no scenario in which crimes will disappear from our lives. For this reason, we took the crimes history and developed a model that aims to predict crime rates and help decrease the level of the crimes in Chicago. In order to estimating the crime rates, we have used Gaussian process and during the implementation of the algorithm we inspected few covariance functions and figure out which of the covariance functions provides more accurate results.



Alissa Pump 11/12/2012
Source: City of Chicago Data Portal
Projection: SPCS83 Illinois East zone:
(US Survey feet) (Transverse Mercator)
Datum: North American Datum 1983/Illinois East (ft US)

Figure 4: Chicago Community Areas Map

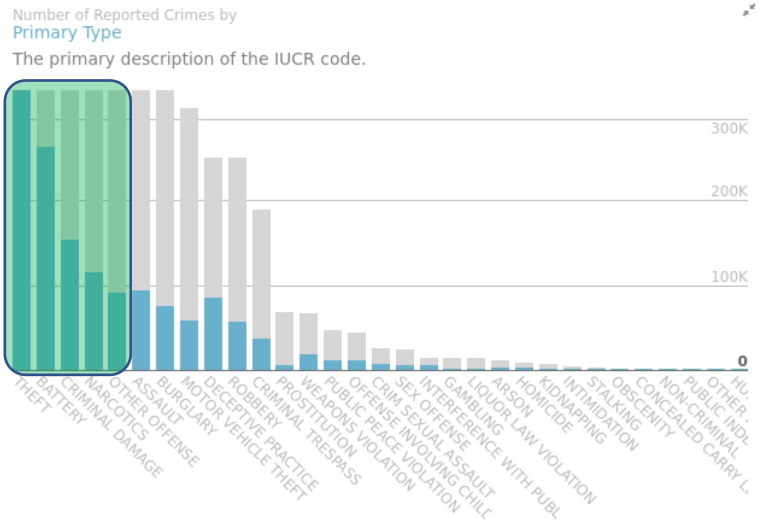
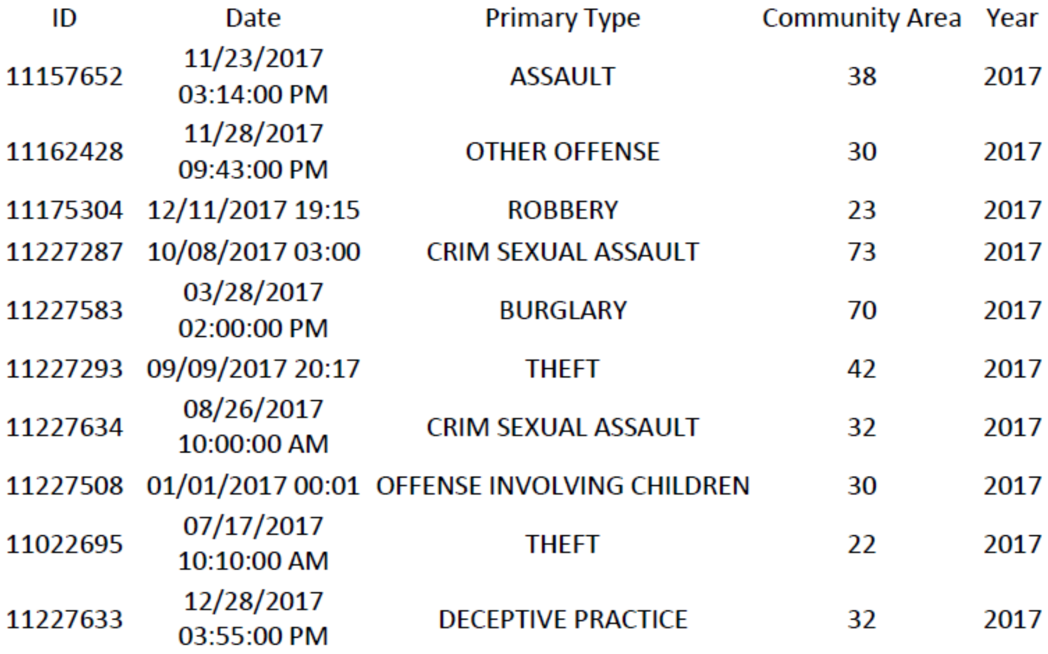


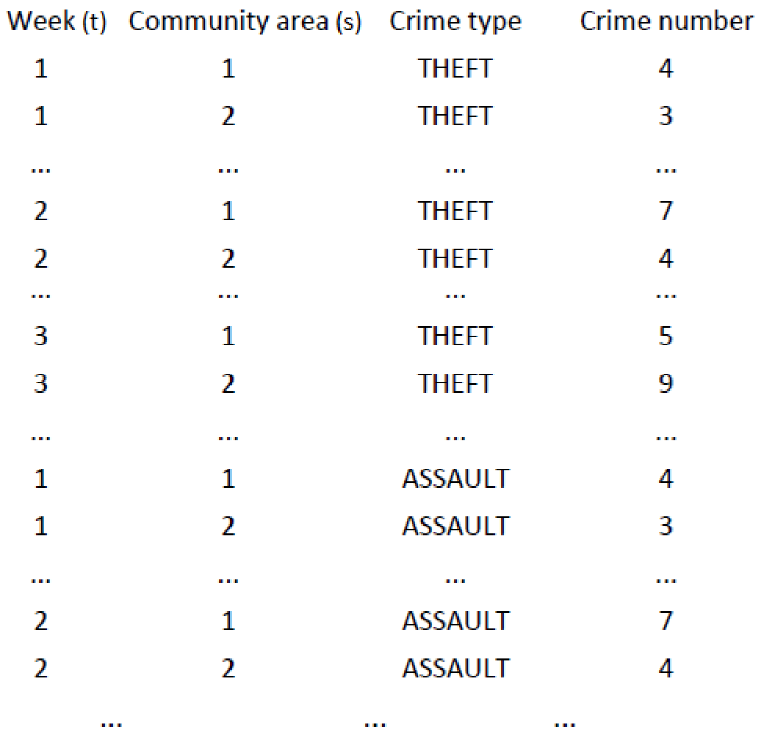
Figure 5: Five common primary types of crimes.

[2] https://data.cityofchicago.org/



ID	Date	Primary Type	Community Area	Year
11157652	11/23/2017 03:14:00 PM	ASSAULT	38	2017
11162428	11/28/2017 09:43:00 PM	OTHER OFFENSE	30	2017
11175304	12/11/2017 19:15	ROBBERY	23	2017
11227287	10/08/2017 03:00	CRIM SEXUAL ASSAULT	73	2017
11227583	03/28/2017 02:00:00 PM	BURGLARY	70	2017
11227293	09/09/2017 20:17	THEFT	42	2017
11227634	08/26/2017 10:00:00 AM	CRIM SEXUAL ASSAULT	32	2017
11227508	01/01/2017 00:01	OFFENSE INVOLVING CHILDREN	30	2017
11022695	07/17/2017 10:10:00 AM	THEFT	22	2017
11227633	12/28/2017 03:55:00 PM	DECEPTIVE PRACTICE	32	2017

Table 1: Database from the website CityOfChicago



Week (t)	Community area (s)	Crime type	Crime number
1	1	THEFT	4
1	2	THEFT	3
...	...	...	...
2	1	THEFT	7
2	2	THEFT	4
...	...	...	...
3	1	THEFT	5
3	2	THEFT	9
...	...	...	...
1	1	ASSAULT	4
1	2	ASSAULT	3
...	...	...	...
2	1	ASSAULT	7
2	2	ASSAULT	4
...	...	...	

Table 2: Database table aggregated by neighborhood and week

4. PRELIMINARY SOFTWARE ENGINEERING DOCUMENTS

4.1 Flow Diagram

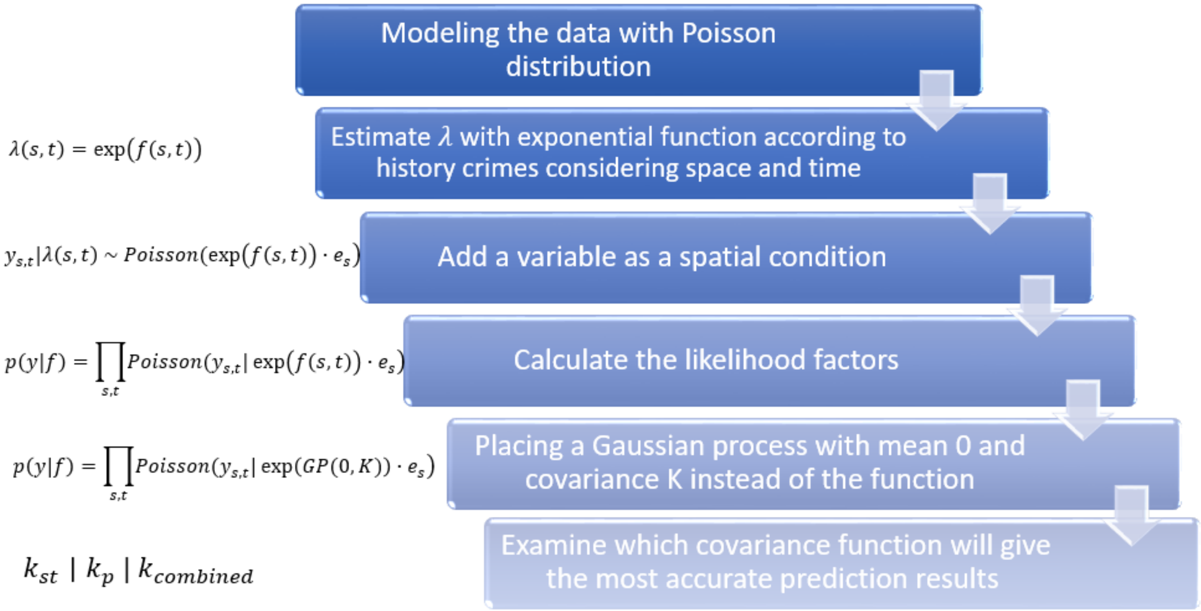

Modeling the data with Poisson distribution
$\lambda(s,t) = \exp(f(s,t))$
Estimate λ with exponential function according to history crimes considering space and time
$y_{s,t}|\lambda(s,t) \sim Poisson(\exp(f(s,t)) \cdot e_s)$
Add a variable as a spatial condition
$p(y|f) = \prod_{s,t} Poisson(y_{s,t}| \exp(f(s,t)) \cdot e_s)$
Calculate the likelihood factors
$p(y|f) = \prod_{s,t} Poisson(y_{s,t}| \exp(GP(0,K)) \cdot e_s)$
Placing a Gaussian process with mean 0 and covariance K instead of the function
$k_{st} \mid k_p \mid k_{combined}$
Examine which covariance function will give the most accurate prediction results

Figure 6: Algorithm Flow Diagram

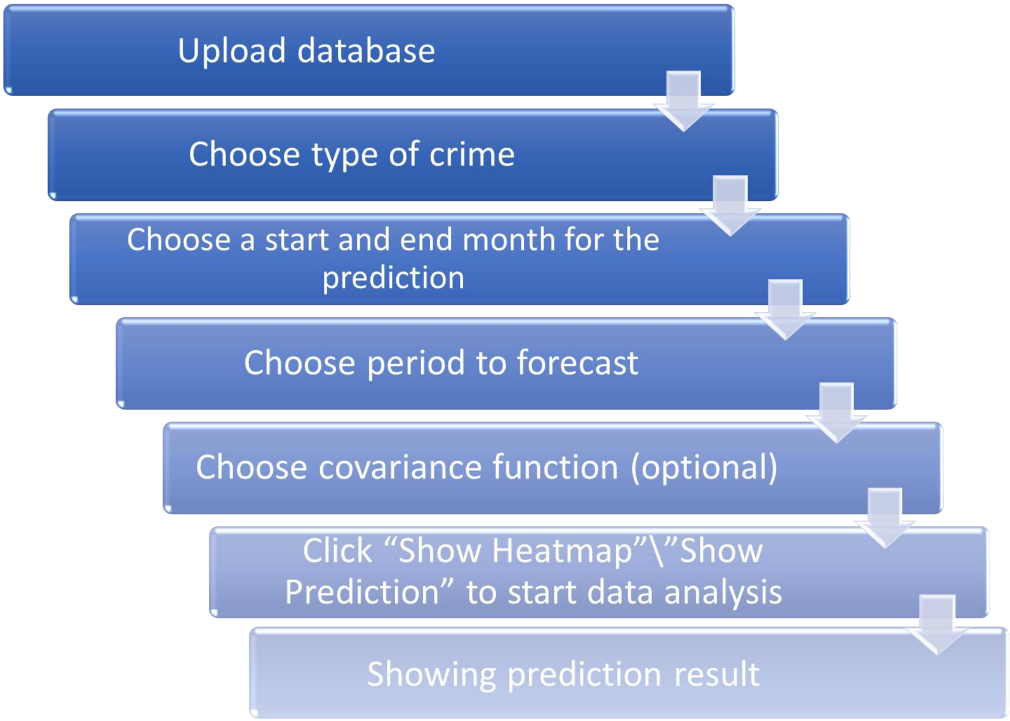

Upload database
Choose type of crime
Choose a start and end month for the prediction
Choose period to forecast
Choose covariance function (optional)
Click “Show Heatmap”\”Show Prediction” to start data analysis
Showing prediction result

Figure 7: GUI Flow Diagram

4.2 Requirements (Use Case)

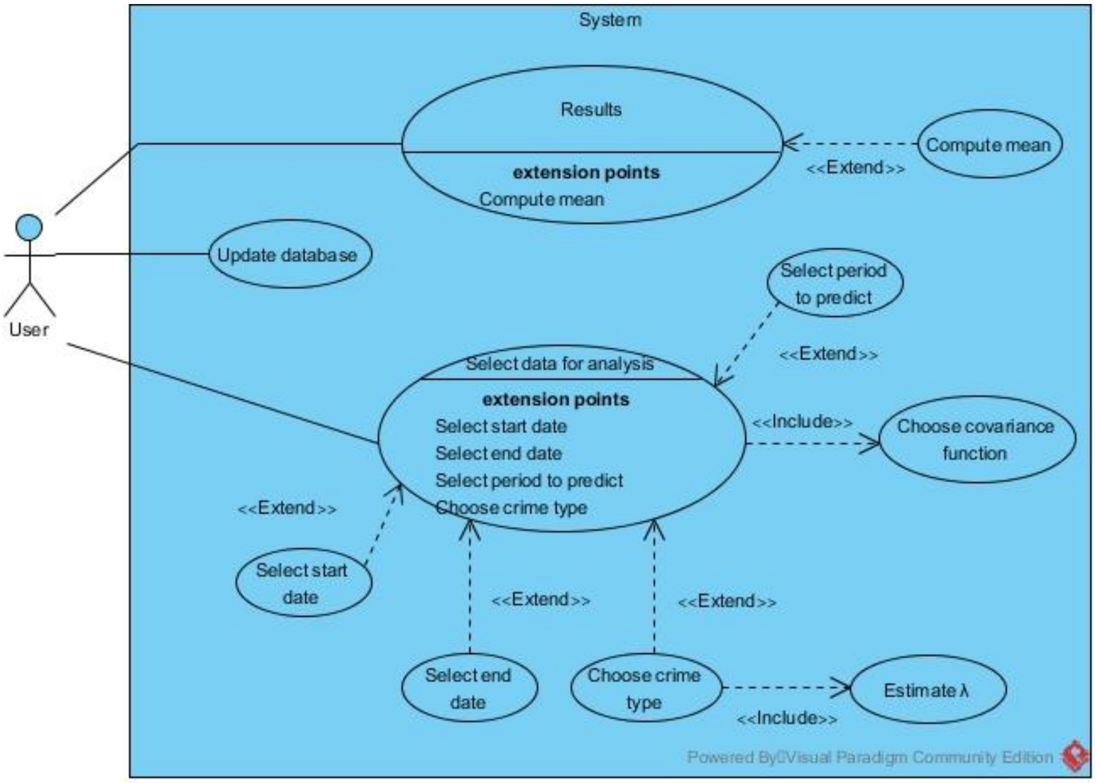


Figure 8: Use Case Diagram

4.3 Design (GUI, UML Diagrams)

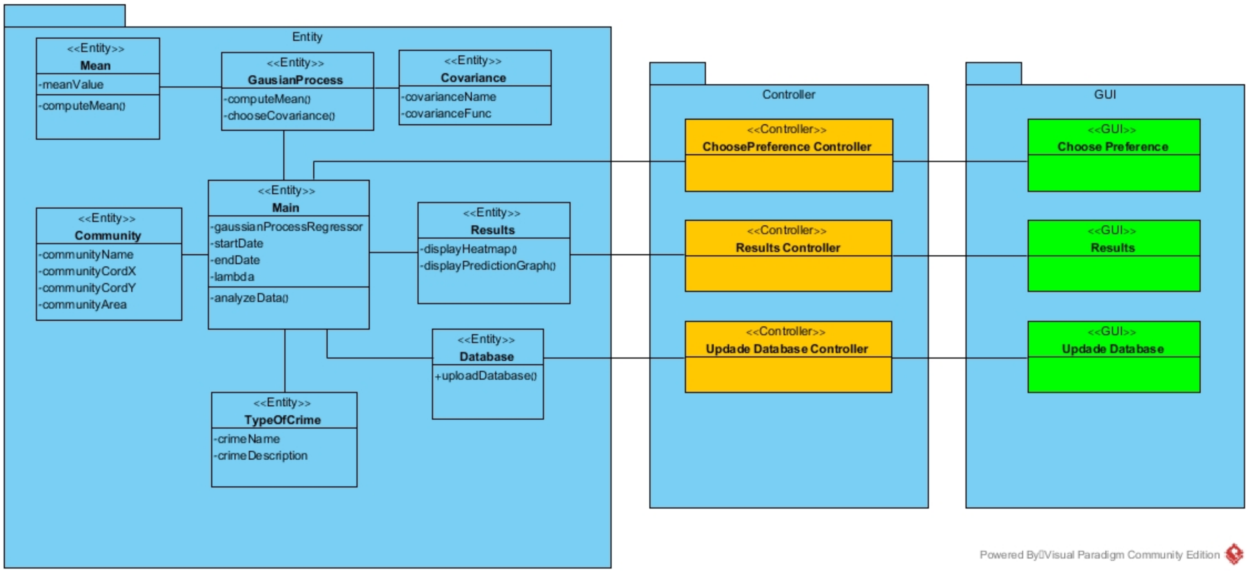
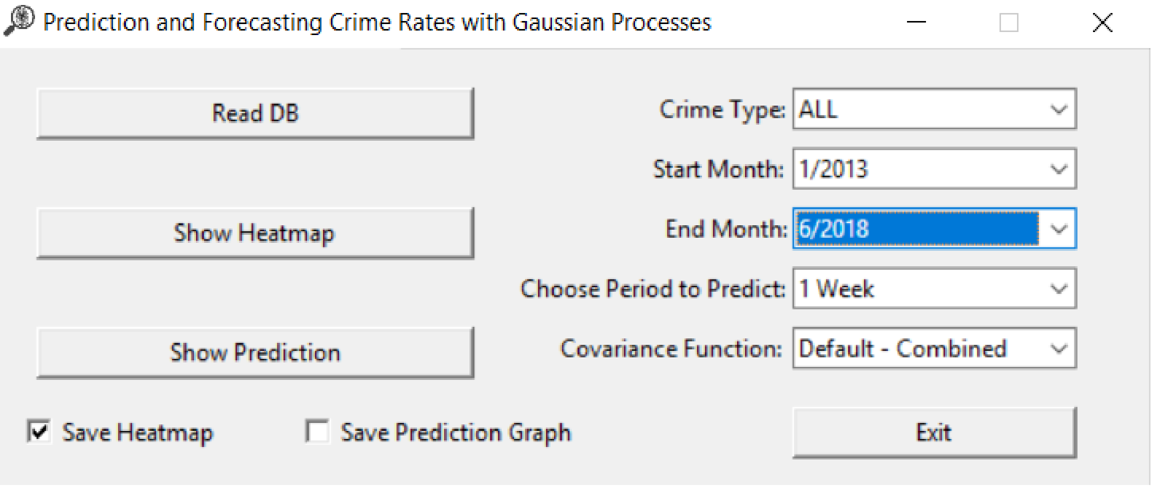


Figure 9: Class Diagram

4.3.1 GUI and User manual


Prediction and Forecasting Crime Rates with Gaussian Processes
Read DB
Show Heatmap
Show Prediction
Exit

Crime Type: ALL

Start Month: 1/2013

End Month: 6/2018

Choose Period to Predict: 1 Week

Covariance Function: Default - Combined

☑ Save Heatmap □ Save Prediction Graph

Figure 10: Main GUI, choose preference to analyze

Main Window:

Read DB button - Use load the updated database. Load DB before using the other options.

After choosing all the preferences of the prediction in the combobox use this buttons:
Show Heatmap button - show the prediction results on heatmap of Chicago.
Show Prediction button - show the prediction results on graph.
Save Heatmap/Save Prediction Graph - use this checkboxes to save the results as pictures in folder Image at the project folder.

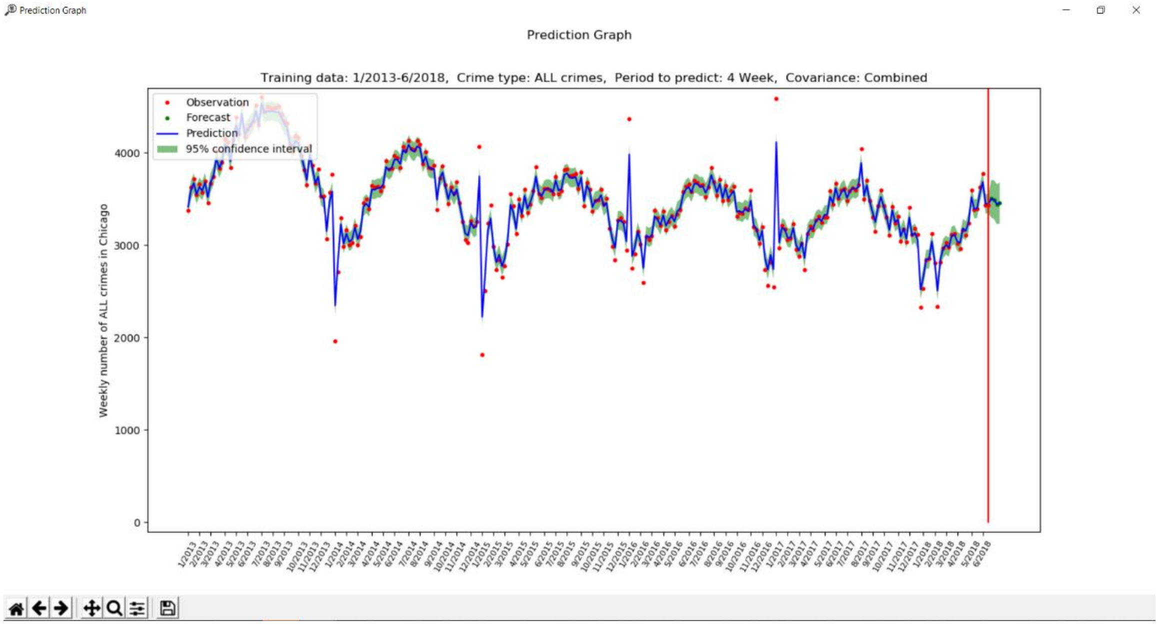
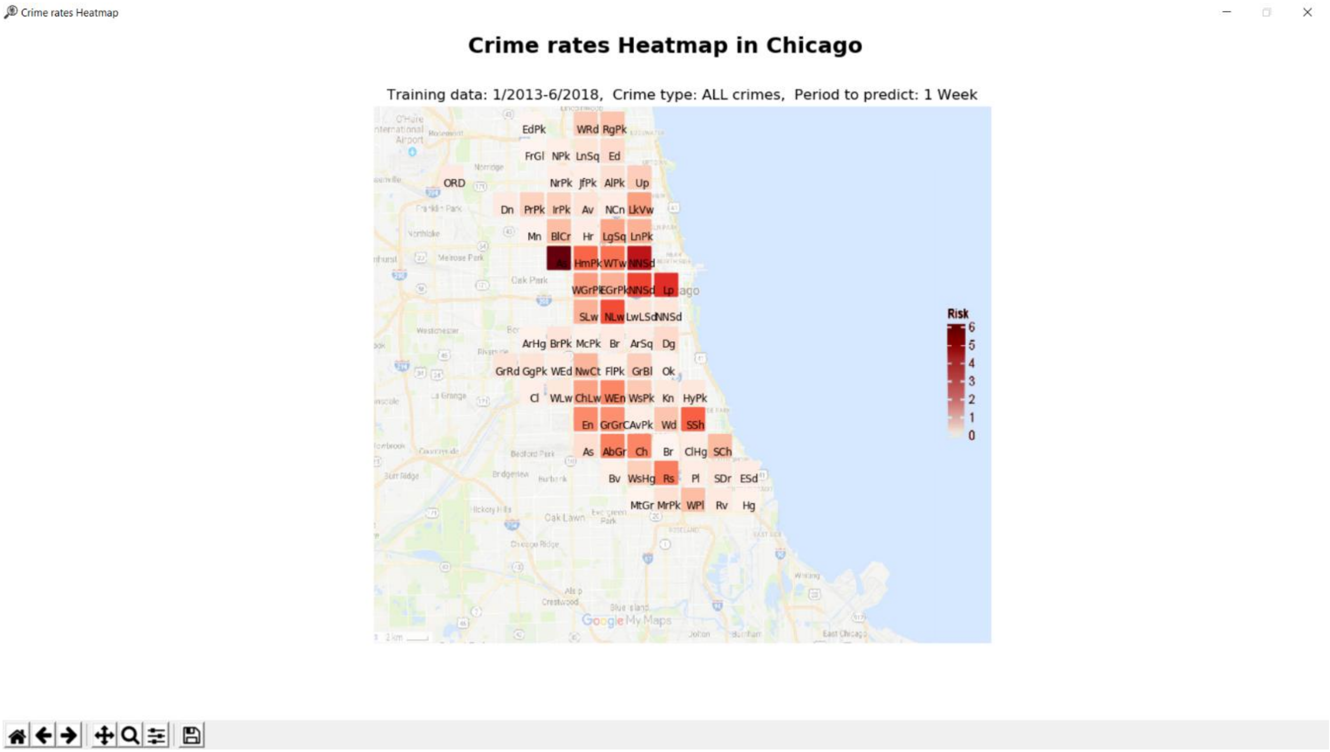


Figure 11: GUI of the Prediction

Prediction Graph:

All the selected preferences show above the graph. Axis X represent the week in the year, and axis Y represent the weekly number of crimes in Chicago. The red dots are the observations of the crime rates. The forecast is to the left of the red vertical line. The green dots are the forecast of the amount of crime rates. The blue line represents the prediction line that the Gaussian process generated. 95% uncertainty intervals are shown in green.



Crime rates Heatmap

Crime rates Heatmap in Chicago

Training data: 1/2013-6/2018, Crime type: ALL crimes, Period to predict: 1 Week

Figure 12: GUI of the Heatmap, representing the prediction results

Heatmap:

All the selected preferences show above the heatmap.
Each square represent community area, and the intense of the color representing the risk of the crimes amount in every one of them.

4.4 Testing Plan

In order to check out the system performance we will run the program on some significant input:

Test No.	Test subject	Expected result	Actual results
1.	Uploading incorrect file to database.	An error message appears: "Please select correct database to read."	Pass
2.	Choose End date that earlier than Start date.	An error message appears: "Please select dates correctly."	Pass
3.	Press "Show Heatmap" at main screen before choosing preferences.	An error message appears: "Please insert the database."	Pass
4.	Press "Show Prediction" at main screen before choosing preferences.	An error message appears: "Please insert the database."	Pass
5.	Press "Show Heatmap" at main screen before uploading a database.	An error message appears: "Please select correct database to read"	Pass
6.	Press "Show Prediction " at main screen before uploading a database.	An error message appears: "Please select correct database to read"	Pass
7.	Press "Exit" at main screen.	A message appears: "Are you sure you want to exit?"	Pass
8.	Press "Show Heatmap" after choosing the required input.	The system displays the prediction heatmap.	Pass
9.	Press "Show Prediction" after choosing the required input.	The system displays the prediction and forecasting graph.	Pass
10.	Press "Show Heatmap" and "V" on the "Save Heatmap" checkbox.	The system displays the prediction heatmap and save png file at <project folder>\Image	Pass
11.	Press "Show Prediction" and "V" on the "Save Prediction" checkbox.	The system displays the prediction and forecasting graph and save png file at <project folder>\Image	Pass

Table 3: Testing plan

5. RESULTS AND CONCLUSION

5.1 Results

5.1.1 Choosing the covariance function

As we said before using Gaussian process function in the model where the mean is 0 and a various of covariance functions K - $f \sim GP(0, K)$. There are many ways to construct covariance function, and each one of these functions will affect differently on the prediction results.

In our tests we examined these four Covariance functions:

1) k_s - spatial covariance (Figure 13)
2) k_t - temporal covariance (Figure 14)
3) k_p - periodic covariance (Figure 15)
4) $k_{combined}$ - combined covariance (Figure 16)

These are the results of the tests:

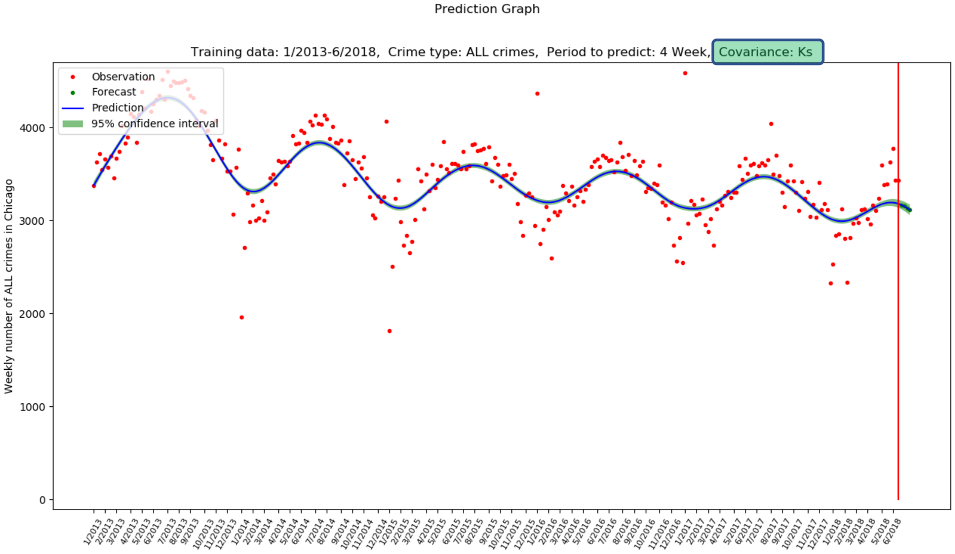


Figure 13: GP function using Covariance function k_s

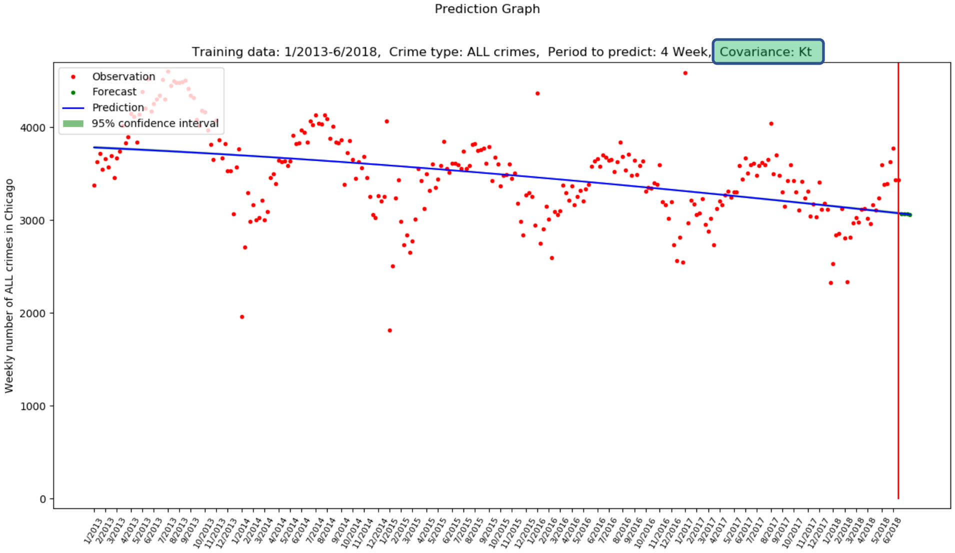


Figure 14: GP function using Covariance function k_t

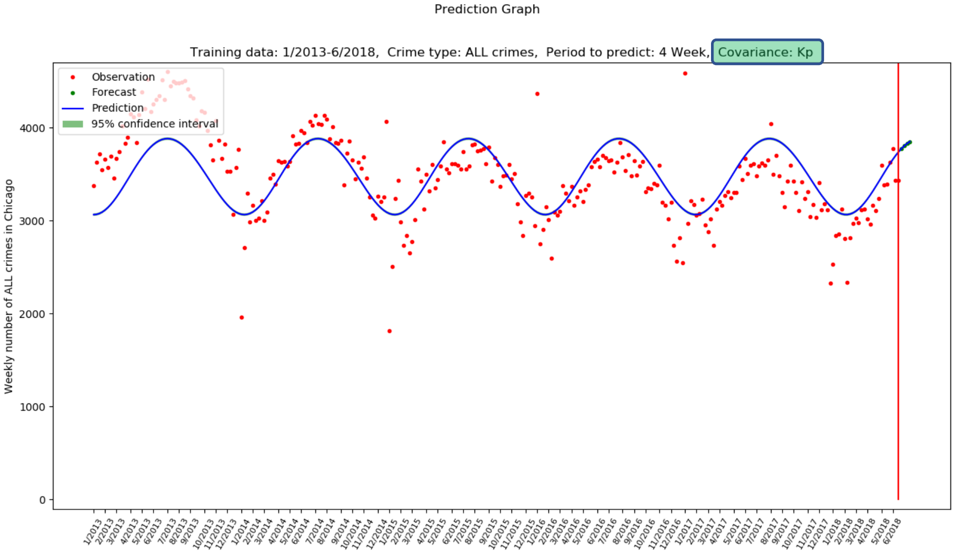


Figure 15: GP function using Covariance function k_p

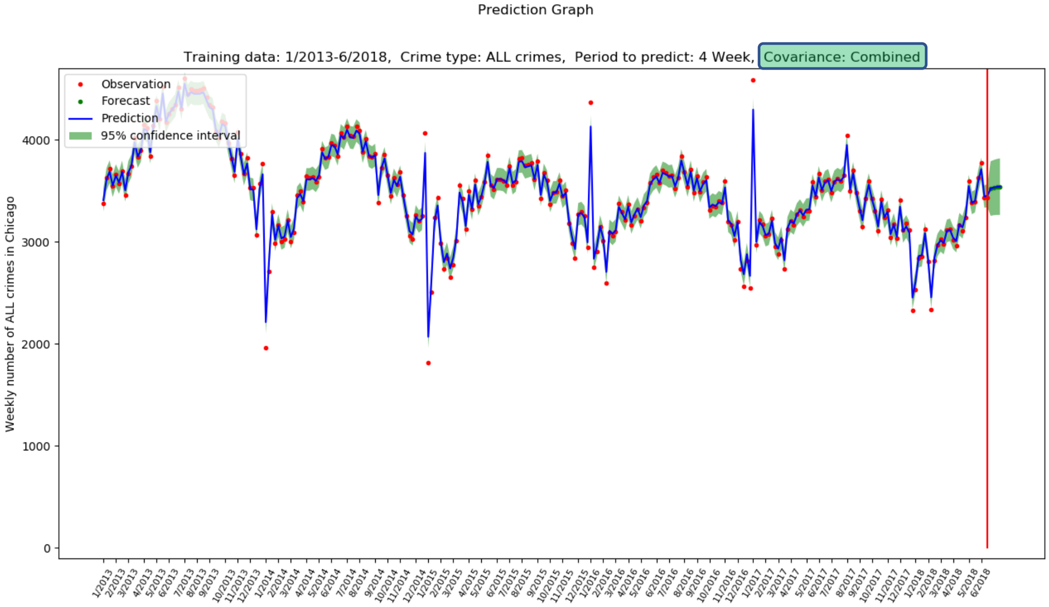


Figure 16: GP function using Covariance function $k_{combined}$

As we can see after running the Gaussian process with several covariances there is one covariance function that provides the most accurate results. While we examined the k_s covariance (Figure 13), it considers only the aspect of the spatial term and we got prediction that is not accurate for all the data of the crime rates. At the examination of the k_t covariance (Figure 14), we saw that because it only considers the aspect of the temporal term, we get prediction result of a linear line that will not be accurate at all if we use it to forecast the crime rates. When we examined the k_p (Figure 15) covariance, we saw the fact that it considers only the aspect of the periodic temporal term, the prediction was with seasonal variations and it cause a circular line that's not representing the real data. The last covariance that we checked is $k_{combined}$ (Figure 16) this covariance considers all the aspects, it contains the combination of the spatial, the temporal and the periodic terms, its prediction gives the closest line to represents the real data and due to that provides the most accurate results in ours examines.

5.1.2 The method of drawing the colors intensity

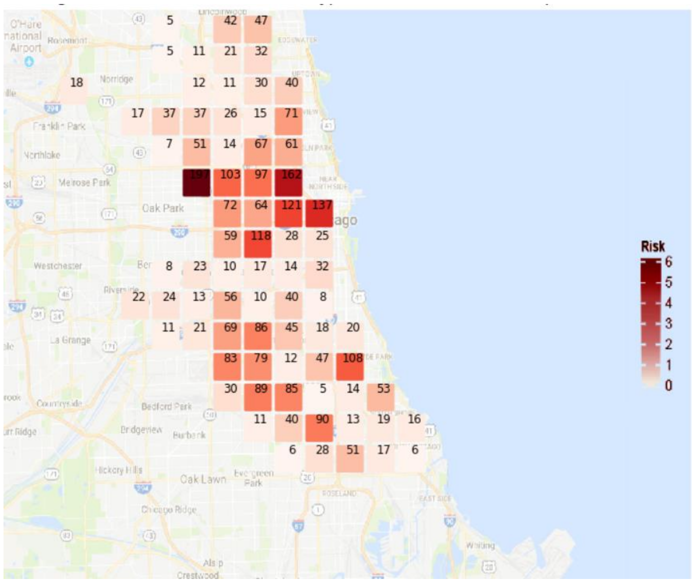


Figure 17: Heatmap with the normalized numbers

On the next step using the DB of Table 2 and the selected kernel K, predict and forecast crime number for each one of the community area c_i, $i = 1, \dots, m$. c_i is forecast crime number of ith community area and m is total number of community area (77). To draw the heatmap, we have normalized these forecast crime numbers as $c_i/\max(c)$ and draw heatmap.

5.1.3 Examination of the Heatmap

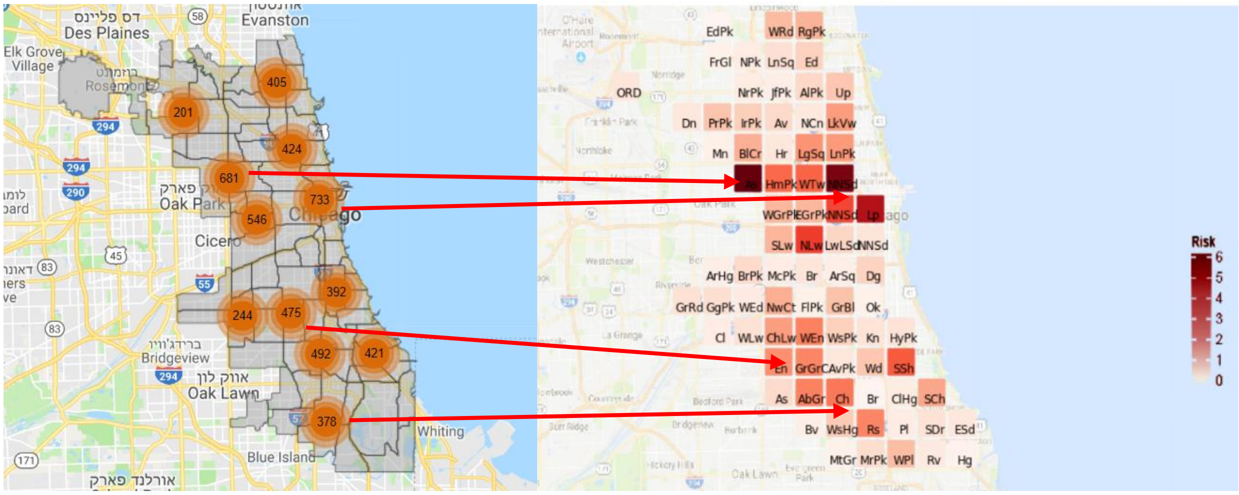


Figure 18: The forecast results versus the real information

In Figure 18, On the right we have the prediction result on heatmap of the first week in the year 2018 based on training period 1/2017-12/2017, and on the left we have the real information from the website "CityOfChicago" of the same dates. When we examined the prediction result we can see that its providing correct crime rates of this dates. So therefore, it confirms from another point of view that the algorithm does works as required.

5.2 Conclusion

We presented a general framework for the statistical modeling of spatiotemporal count data. Developing an algorithm of prediction and forecasting crime rates, using stochastic process called Gaussian process. This system is developed and designed to serve law enforcement agencies, but it is quite generic and can be easily be transformed to other spatiotemporal data usage, such as weather forecasting or soccer match results. In order to being able to predict crimes the system uses an extensive database of committed crimes that were established and analyzed for that purpose. Our software has one weakness, when it runs with small amount of data as a base for the prediction, the forecasting won't be accurate. To overcome that weakness, after loading the database the preferences of the dates and covariance for the prediction has a default value (it selects the whole database as training data, and the combined covariance function) and we recommended not to change these settings to get optimal results. This prediction and forecasting will help the police to identify places and times with the highest risk of crime, and then optimize the positions of the units in the field by sending them to the dangerous location according to the forecasting, thereby preventing crime before it occurs.

6. REFERENCES

[1] Seth R. Flaxman, "*A General Approach to Prediction and Forecasting Crime Rates with Gaussian Processes*", 2014.

[2] A. Cohen, and R.H. Jones, "Regression on a random field", Journal of the American Statistical Association, 64:1172-1182, 1969.

[3] Carl Edward Rasmussen and Christopher K. I. Williams, The MIT Press, "Gaussian Processes for Machine Learning", 2006.

[4] C. Holmes, B. Mallick, and H. Kim, "Analyzing non-stationary spatial data using piecewise Gaussian processes", technical report presented at Bayesian Statistics 7, 3:4-15, 2002.

[5] C.J. Paciorek. "Nonstationary Gaussian Processes for Regression and Spatial Modelling" PhD thesis, Carnegie Mellon University, Pittsburgh, Pennsylvania, 5:70-85, 2003.

[6] D.J.C. MacKay. "Introduction to Gaussian processes". Technical report, Univ. of Cambridge, 13-20, 1997.

[7] K.V. Mardia, and C.R. Goodall, "Spatial-temporal analysis of multivariate environmental monitoring data". In: Multivariate Environmental Statistics 6, New York, 16:348-365, 1992.

[8] L. Anselin, J. Cohen, D. Cook, W. Gorr, and G. Tita. "Spatial analyses of crime"', 4:221–225, 2000

[9] P.D. Sampson, D. Damian, and P. Guttorp, "Advance in modeling and inference for environmental processes with nonstationary spatial covariance", National Research Center for Statistics and the Environment University of Washington 6-14, 2001.