A THROUGH SPATIO-TEMPORAL ANALYZER VIA R PROGRAMMING ENVIRONMENT INTEGRATED WITH COMPARATIVE PREDICTION ANALYSIS

1. INTRODUCTION

A spatio-temporal object can be defined as an object that has at least one spatial and one temporal property. The spatial properties are location and geometry of the object. The temporal property is timestamp or time interval for which the object is valid [1]. Spatiotemporal datasets are used to observe spatial changes of objects' different attributes over time. There are significant amounts of spatio-temporal datasets in many different application domains like; meteorology, medicine, transportation and biology. Some example spatio-temporal datasets are satellite images of parts of the earth, temperature readings for a number of nearby stations, election results for voting districts and a number of consecutive elections, trajectories for people or animals possibly with additional sensor readings, disease outbreaks or volcano eruptions [2]. Considering the fact that spatiotemporal datasets are found in many different application domains and usually very large, they contain a lot of information and have important role in decision support. Hence, analysis and visualization tools to extract and understand this significant patterns and information are needed.

Spatiotemporal analysis can be categorized as temporal data analysis, spatial data analysis, dynamic spatiotemporal data analysis and static spatiotemporal data analysis. Wind speed measurements in a given area, and the analysis of autocorrelation between measurements in different time points can be an example of temporal analysis in which the spatial properties are considered constant. Analysis of the change of temperature while moving away from seacoast is an example of spatial analysis which time component is constant. Analysis of traffic density over time in an area or moving storms is categorized as dynamic spatio-temporal analysis[1].

Spatio-temporal data mining is a user-centric, interactive process, where both mining experts and domain experts work closely together to gain insight on the spatial and temporal evolution of a given phenomenon [3]. It is important to analyze both spatial and temporal properties, in order to understand attribute relations, make predictions and interpret the nature of the spatio-temporal object. Visualization and data mining techniques are widely recognized to be powerful in this domain [4], they are very useful in transforming the large, complex data into easy to interpret summaries of different attributes, like mapping a large crime data to locate the safest neighbourhoods of a city or finding the correlation between ethnicity populations and crime rates. However, visualization features and analysis tools in many existing applications are limited, and usually are designed to analyze either the spatial or temporal aspects of objects alone. Hence, a new user-friendly spatio-temporal analysis and visualization application is needed to be developed.

In the present paper, a graphical user interface developed in RStudio is presented. The aim of this system is to combine different spatio-temporal analyzing tools from different R packages in a user-friendly interface, which will allow users to upload their datasets, group and analyze chosen attributes, obtain map visualizations, observe spatial and temporal correlations and a variety of different plots. The application is very easy to use and allows the user to

manipulate the data and change parameters dynamically in many different ways in order to see the patterns and interpret relations.
The proposed application is tested on two real-world datasets, the first dataset is an extended version of San Francisco Crime between 2003-2015 (see [5]), used for analyzing the different type of crimes rates over years, day of the week, hour of the day, to locate safe districts, to observe correlation between different attributes such as correlation between single population and crime rate etc. The second dataset is the Cigar dataset (see [6]), which is considerably a smaller dataset, is used for analyzing the cigarette price changes between states, correlation between sales of cigarette packs per inhabitant and producing price etc.
The remainder of the paper is structured as follows Section 2 explains the related works on spatial, temporal and spatio-temporal data analysis, and the packages used in the application. Section 3 provides detailed information about the description of the tool, and the studied cases while testing the application. Finally Section 4 presents conclusions and ideas for future work.

3. DESCRIPTION OF THE TOOL

3.1. Application Structure

The application structure has seven different conditional panels which separates analysis functions and visualization features obtained from different R packages. The panels are designed to offer easy manipulation of data so that anyone could use the graphical user interface to analyze their spatio-temporal dataset without needing any R code background. In order to use the application effectively and to obtain accurate results and visuals the uploaded dataset should be in long format which can be seen in Figure 1, and for mapping with polygon fillings a shape file containing .shp, .shx and .dbf documents should be uploaded at the same time. Other filtering, grouping and subset options are provided in each panel. The conditional panels provided in the user interface are a basic data frame observing panel, Location Plot panel, Mapping panel, St Plot panel, Moran Plot panel, Correlation Plot panel, and a panel for simple ggplots.

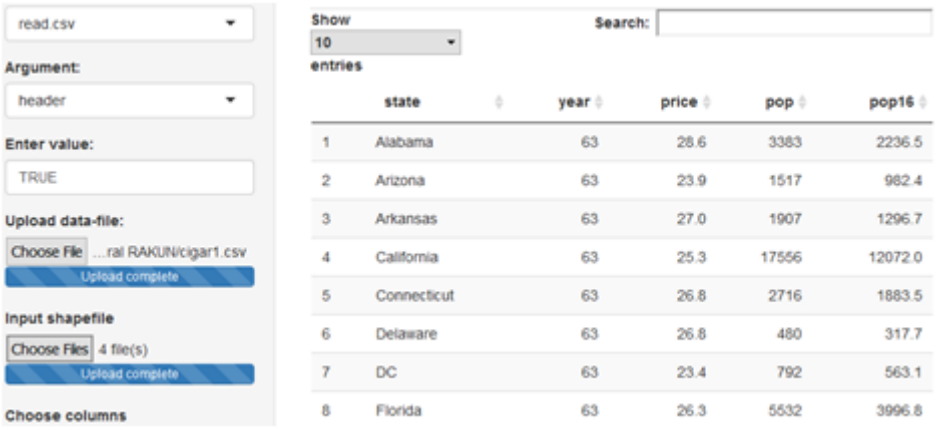
	state	year	pcap	hwy	water	util	pc	gsp	emp
1	ALABAMA	1970	15032.67	7325.80	1655.68	6051.20	35793.80	28418	1010.5
2	ALABAMA	1971	15501.94	7525.94	1721.02	6254.98	37299.91	29375	1021.9
3	ALABAMA	1972	15972.41	7765.42	1764.75	6442.23	38670.30	31303	1072.3
4	ALABAMA	1973	16406.26	7907.66	1742.41	6756.19	40084.01	33430	1135.5
5	ALABAMA	1974	16762.67	8025.52	1734.85	7002.29	42057.31	33749	1169.8

Figure 1. Long Format Dataset

3.1.1. Dataset Panel

The dataset panel is the data uploading and viewing page, before moving on to other panels to analyze data user needs to upload the long formatted dataset and if necessary the shape files. Datasets can be saved in many different formats such as csv, csv2, table, txt etc.

Considering the fact that spatio-temporal datasets are often very large in this panel the data can be displayed based on only chosen columns and can be filtered by using the search bar.



	state	year	price	pop	pop16
1	Alabama	63	28.6	3383	2236.5
2	Arizona	63	23.9	1517	982.4
3	Arkansas	63	27.0	1907	1296.7
4	California	63	25.3	17556	12072.0
5	Connecticut	63	26.8	2716	1883.5
6	Delaware	63	26.8	480	317.7
7	DC	63	23.4	792	563.1
8	Florida	63	26.3	5532	3996.8

Figure 2. Dataset Panel

An example representation of the panel can be seen in Figure 2 above.

3.1.2. Location Plot Panel

The location plot panel enables users to plot data points on user defined maps. In order to use this feature dataset should contain latitude and longitude values such as the lat. and long. values of the location where a crime is committed. In this panel 'get map( )' function from the R "ggmap package" [8] is used to obtain the required map. 'get map ()' function obtains maps from online sources like "Google Maps" or "Stamen Maps". The user has two options the get the map, first option is to enter left/bottom/right/top coordinates of the area, the second option is to enter the location name like "San Francisco", however the first option provides more detailed and exact maps.

Figure 3. Crimes in San Francisco Between 2003 and 2015

The plot of crimes committed between 2003-2015 in San Francisco map can be seen in Figure 3 above.

3.1.3. Mapping Panel

The mapping panel enables users to make ggplot that can fill the studied area's map which is divided in polygons representing districts, counties, states etc. based on the chosen attributes values. The panel uses R "ggplot2 package" [9] to make plots, fortify shape file to merge Spatial Polygons Data Frame (shape file) with the dataset and "rgdal package" [10] to read the shape files in the first place. In order to make sure the locations in the dataset are matched with the polygons in the spatial polygons data frame, two datasets are merged by location IDs or location names at the beginning. After data are merged and the attribute is chosen user can visualize the map by choosing the desired time component.

3.1.4. St Plot Panel

St plot function in the space-time R package [11] uses objects deriving from St class to create trellis plots. St plot panel uses the STFDF function to merge and transform the spatial data frame and dataset into a spatio-temporal object. POSIXct date-time conversion function is used to create the time object needed in STFDF function. As in the other panels, St plot panel has an option to filter the data which is in some cases necessary to transform the dataset to an St object such as crime datasets. An example St plot of cigarette sales of cigarette packs per inhabitant in each state of USA between years 1962 and 1992 can be seen in Figure 4 below.

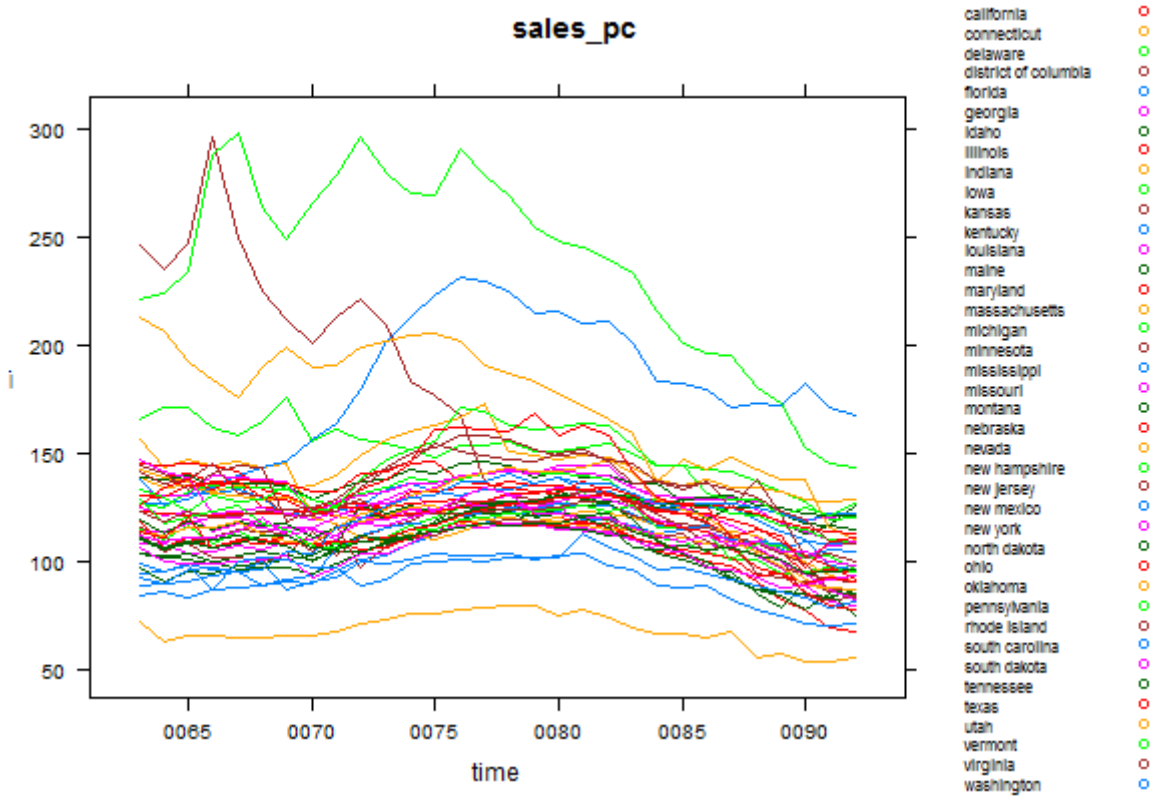


Figure 4. St plot Cigar Sales Prices Between 1962 and 1992

3.1.5. Plot Panel

Plot panel is used for making simple ggplots to observe numeric variables' relations with each other. It uses the ggplot function in ggplot2 package [9], lets the user choose which variables to plot on X and Y axis. The panel has an option to filter the dataset as in the other panels, filter data button in all panels uses filter( ) , group_by_( ), summarize( ), functions from dplyr R package [12] to group data by chosen attributes or filter by searched words. An example ggplot of "Crimes Committed Close Distance to a Police District between 2003-2015" can be seen in Figure 5 below.

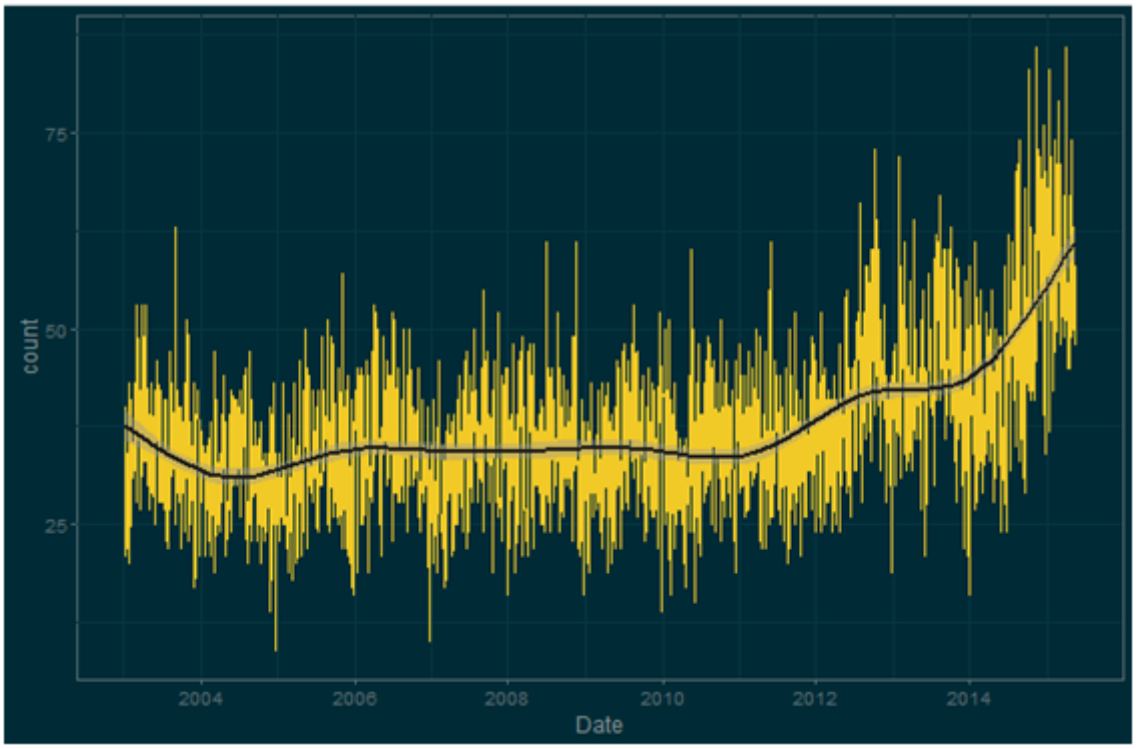


Figure 5. Crimes Committed Close by A Police District

3.1.6. Moran Plot Panel

Moran plot panel offers users to observe if there is spatial autocorrelation between the values of a chosen attribute. Spatial autocorrelation measures the correlation of a variable with itself through space. Spatial autocorrelation can be positive or negative. Positive spatial autocorrelation occurs when similar values occur near one another. Negative spatial autocorrelation occurs when dissimilar values occur near one another. [13] Moran plot panel uses spdep R package's [14] poly2nb ( ) function to obtain neighborhoods from spatial polygons data frame, nb2list ( ) function to obtain the listw object, and moran.plot function to make the plot. The plot gives the attributes values against its spatially lagged values. Hence, the moran plot enables users to assess how similar an observed attribute value to its value obtained from neighboring observations, higher similarity shows high spatial autocorrelation. An example moran plot of Cigarette prices in US states in 1963 can be seen in Figure 6 below.

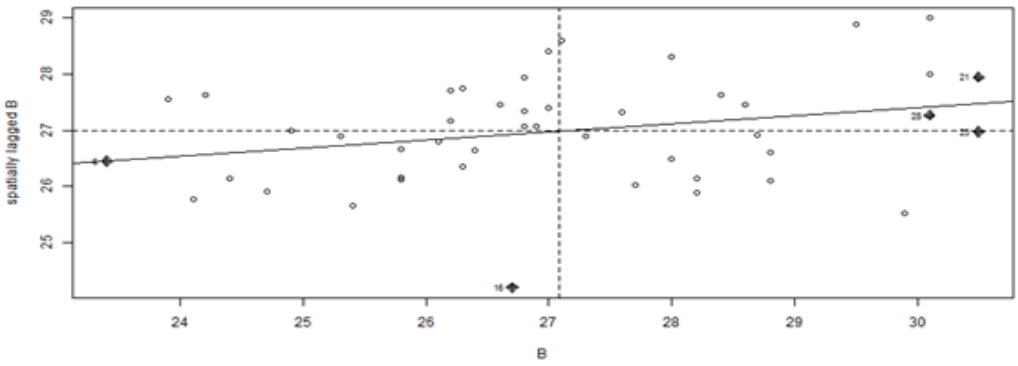


Figure 6. Cigarette Price Moran Plot

3.1.7. Correlation Plot Panel

Correlation plot panel is used for observing the correlations between attributes to interpret their relations. The panel uses the corrplot R package's [15] corrplot( ) function to display correlations of variables between [-1,1] in a matrix with circle method. The example of counted crimes with arrest resolution in San Francisco with ethnic populations can be seen in Figure 7 below.

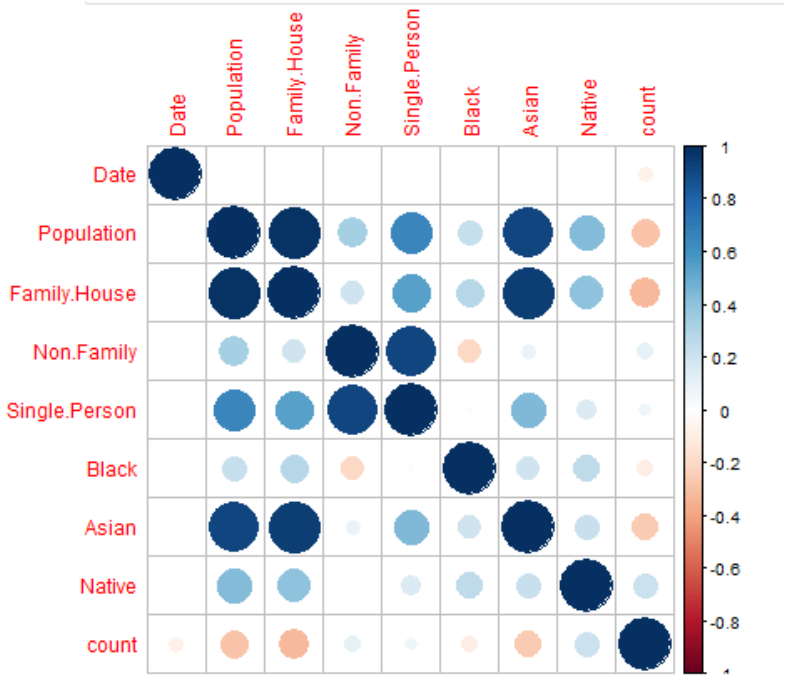


Figure 7. Correlation Between Ethnic Populations and Crime

3.2. Datasets

The application's structure is developed by using different R packages' different functions as described in the section above. All the packages and functions that are used in the GUI can be found in appendix A. After the base structure is built, different datasets with different attributes and sizes were needed to test the panels' functions, and to verify if any other

features are needed to complete the work. To this aim, two datasets were selected. The first dataset contains annual cigarette consumption in 46 US states between 1963 and 1992, stored in long format. Cigar dataset is a simple spatio-temporal dataset with only a few attributes, year and location ID information. The second dataset is an extended version of San Francisco Crime dataset which is one of the most analyzed and used datasets to work on spatio-temporal analysis and visualization techniques. Both datasets are used for testing each panel's functions. Tests showed that in each panel, a filter data option is needed, and also to ensure locations in shape-file and datasets are matched a prior merging by id feature is added.

3.2.1. Cigar Dataset

Cigar data concerns annual cigarette consumption in 46 US states 1963-1992, and contains columns with year and location IDs, and population, cpi (consumer price index), ndi (net disposable income), price, sales of cigarette packs per inhabitant (sales_pc) attributes. It is analyzed and used for testing all the panels except the location plot panel, since it does not contain any latitude and longitude values. Cigar dataset does not need prior filtering since it is not a very large detailed data and it can be transformed into a spatio-temporal object from its original structure.

When cigar dataset is analyzed using the app, it has verified that all panels are functioning and visuals help interpret the dataset easily. Some example results obtained from the panels are:

- Correlation panel cited that while price has a high positive correlation with cpi and ndi, it has a negative correlation with sales of cigarette packs per inhabitant as expected. Correlation plot can be seen in figure 8 below.

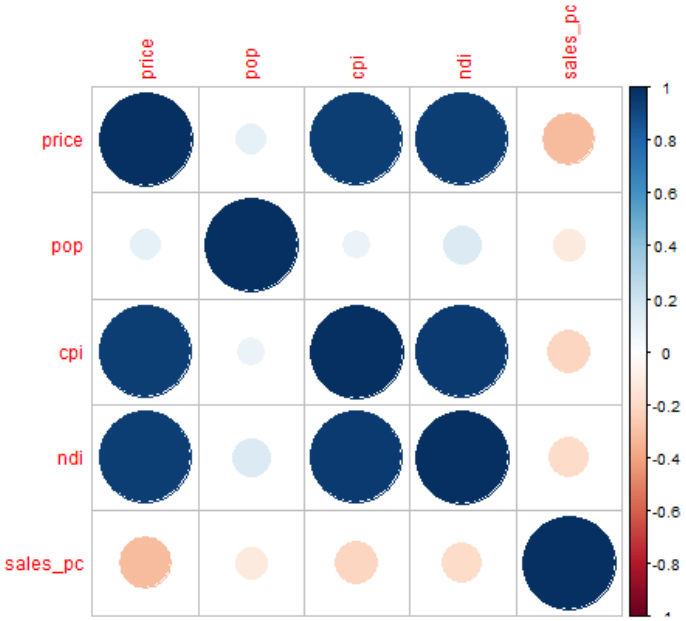


Figure 8. Correlation Plot Cigarette Attributes

- St plot panel and map plot panel displayed that while prices are increasing and decreasing together in all states, Colorado has slightly lower prices over the years, sales of cigarette pack numbers are highest in New Hampshire.
- When attributes are analyzed in moran plot panel, it is observed that spatial autocorrelation exists, and attribute values are very similar to their spatially lagged values.

- Plot panel is also tested with cigar dataset, and verified. Hence the dataset showed that app is functioning without any problem with a simple dataset.

3.2.2. San Francisco Crime Dataset

San Francisco Crime dataset is originally downloaded from sf-open-data platform between 2003 and 2015. The original dataset contains date-time information columns, latitude and longitude values of each committed crime, police districts, category, descript, resolution, address, and location columns. To extend the dataset for further analysis, such as crime rates correlation with population or with distance from the nearest police district etc. new columns are added to the dataset. Since the crime dataset is very large cannot be converted into a spatio-temporal object directly or analyzed easily, filter data function in panels are used to analyze and interpret the dataset.
When San Francisco Crime Dataset is analyzed using the app, it has verified that all panels are functioning and visuals help interpret the dataset easily. Some example results obtained from the panels are:

- Location plot showed that there is a large concentration in Central, Tenderloin and Southern Police Districts. The plot is given in Figure 3 above in section 2.1.2.
- The map plot panel and st plot panel cited that the majority of crimes with all resolution types, and the majority of crimes that are committed close by a police department occur in Southern Police District. The lowest numbers are observed in Mission Police District. The plots can be seen in Figure 9 and 10 below.

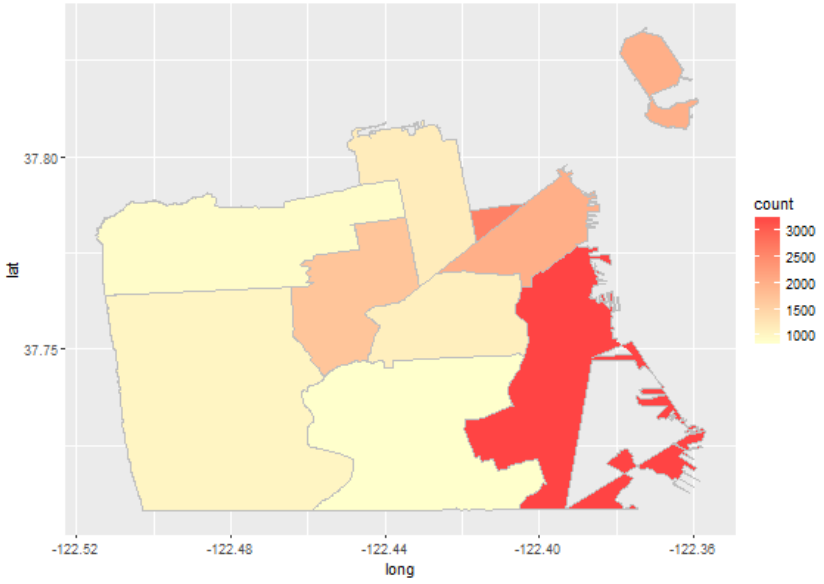


Figure 9. Map Plot of Crime Numbers in Police Districts in 2003

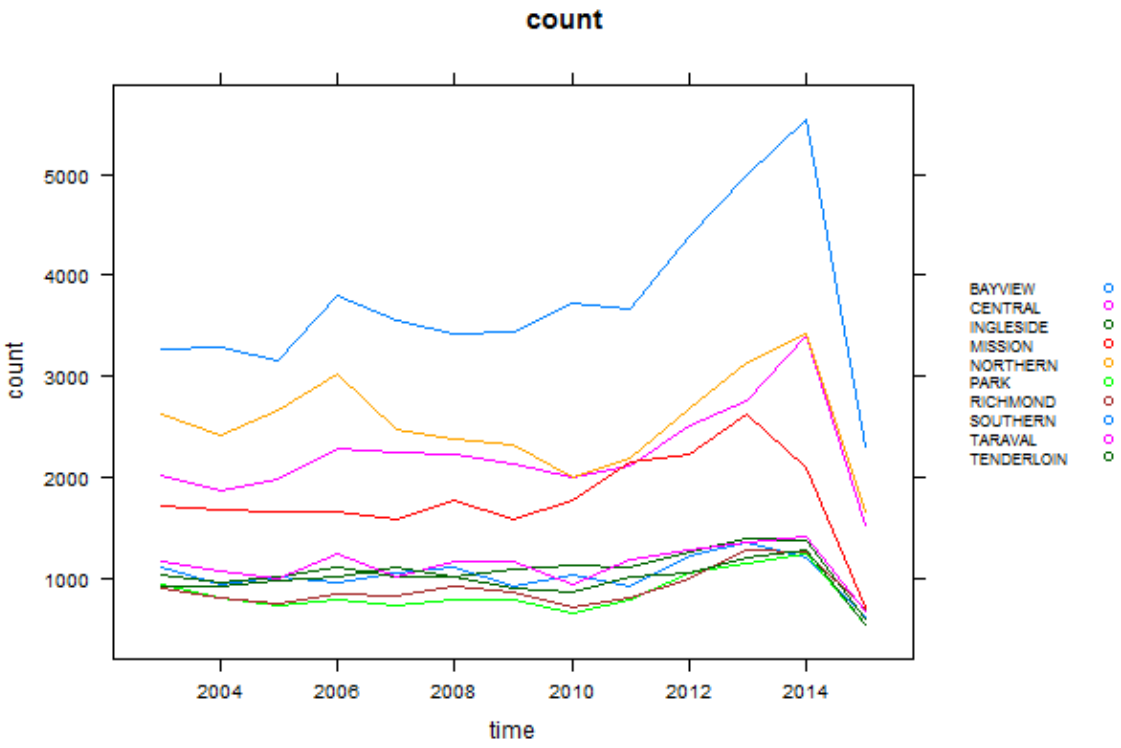


Figure 10. St Plot of Crime Numbers in Police Districts

- When crime rate is analyzed using the moran plot panel, it is observed that there is very low spatial autocorrelation and values are not close to their spatially lagged values.
- Correlation plot panel showed that ethnic population does not have any significant effect on crime rate.
- San Francisco Dataset verified that all panels and filter data functions are working, and the app can be used for easy and fast analysis of complicated, large datasets like crime data.

Package	**Url**	**Function**	**Purpose**	**Developer**
Spacetime[11]	cran.r-project.org/web/packages/spacetime	st_plot(),STFDF(),as.POSIXct()	To create and plot a spatiotemporal dataframe.	Edzer Pebesma
corrplot[15]	cran.r-project.org/web/packages/corrplot	corrplot()	To calculate correlations between given variables and get correlation plot.	Taiyun Wei Viliam Simko
lubridate[16]	cran.r-project.org/web/packages/lubridate	mdy()	To change date format	Vitalie Spinu
dplyr[12]	cran.r-project.org/web/packag	group_by(), summarize()	To group and summarize data for different analysis,	Hadley Wickham

	es/dplyr		plots or correlograms.	
ggmap[8]	cran.r-project.org/web/packages/ggmap	get_map()	To get map of the location based on entered location name or coordinate limits.	David Kahle
ggthemes[17]	cran.r-project.org/web/packages/ggthemes	theme_..()	To select different themes for the plots.	Jeffrey B. Arnold
ggplot2[9]	cran.r-project.org/web/packages/ggplot2	ggplot()	To make different types of detailed plots.	Hadley Wickham
RcolorBrewer [19]	cran.r-project.org/web/packages/RcolorBrewer	colorRampPalette()	To get different color schemes for maps and graphs.	Erich Neuwirth
rgdal[10]	cran.r-project.org/web/packages/rgdal	readOGR(),proj4string(),spTransform()	To read in shapefiles, and to get coordinates based on given CRS.	Roger Bivand
reshape2[20]	cran.r-project.org/web/packages/reshape2	dcast()	To turn rows into columns.	Hadley Wickham
spdep[14]	cran.r-project.org/web/packages/spdep	moran.test(), moran.plot()	To test spatial correlations between neighbouring spots, and to make plot that represents spatial correlation.	Roger Bivand
Shiny[18]	cran.r-project.org/web/packages/shiny	…	To build the interactive web application.	Winston Chang
DT[21]	cran.r-project.org/web/packages/DT	DT: : datatable()	To render data objects as HTML tables.	Yihui Xie
Maptools[22]	cran.r-project.org/web/packages/maptools	readShapePoly()	To read data from a polygon shapefile into a SpatialPolygonsData	Roger Bivand

			Frame object.	