**Abstract**

Time-series analysis is a valuable tool for forecasting and discovering trends to predict future events. The challenge occurs in preparing the data and finding the suitable tools to examine and model such data. The data consists a series of data points recorded in time order and it usually measures the aggregated behavior of a group of people or transactions such as stocks or purchases. A good forecasting can help managers or business leaders to determine future goals based on historical patterns.

**Introduction**

The purpose of applying time series method is to extract meaningful information in the data that can be utilized in decision making for future events. A time-series analysis should provide a broad idea of the trend in data, whether or not it is increasing or decreasing over time. Additionally, if the analysis indicates there is a pattern over a period of time, we can use the pattern to predict a different time point.

For this case study, we will be utilizing a data set containing a series of five-minute interval time points and each interval is the number of cars entering the Glendale on ramp for the 101 North freeway in Los Angeles. The on ramp is located close to the Dodgers stadium but it is not the closest compared to other ramps so the game traffic isn’t as heavy. We will create a time-series model in attempt to find possible trends and verify if there is any correlation between the traffic data and a Dodgers game at the stadium.

**Data Set Description**

The original data contains two sets. The first set is a series of traffic data from April 10, 2005 to October 11. Each row in this set is a five-minute mark with the number of cars entering the on ramp for the previous five minutes. In order to improve performance, we reduce the the number of time points by aggregating the five-minute intervals into thirty-minute intervals. For example, if the recorded time is 13:15:00, the number of cars collected is for the times between 13:00:00 to 13:14:59. The second set is the list of Dodgers games in between April 2005 to September 2005. Each row in this data set is a game played at the Dodger stadium with the game date, start time, end time, attendance and outcome.

For the purpose of our analysis, we are going to pick a subset of the variables from the Dodgers data set and they are the game date, start time and end time. We then merge the Dodgers data with the traffic data set and created a new variable called “*has\_time*” that tracks if a game is on during a time interval. Since we are only interested in the time frame when there is potential traffics on the roads due to the baseball game, we decide to calculate *has\_time* such that if a traffic time interval falls in the range of two hours after the start time and two hours after the end time, we will consider this time interval as there is a game in process.

For example, suppose a game start at 13:10:00 and ends at 16:23:00, the interval that should be marked as there is game would be between 15:30:00 and 18:30:00. We first add two hours to the start time and the end time and we get 15:10:00 and 18:23:00. We then find the next closet thirty-minute intervals that would cover this range of time, and they are 15:30:00 and 18:30:00.

The merged data set consists 8400 rows with 5 explanatory variables and 1 response variable.

|  |  |  |  |
| --- | --- | --- | --- |
| **MTCARS DATA SET** | | | |
| **VARIABLE** | **VARIABLE TYPE** | **DESCRIPTION** | **DATA RANGE (max/min)\*\*** |
| traffic\_time | Explanatory | The date and time this data point is recorded | NA |
| cars\_cnt | Explanatory | The number of cars taking the on ramp | 0 / 340 |
| game\_date | Explanatory | Date of a Dodgers game at the Dodgers’ stadium | NA |
| start\_time | Explanatory | Start time of a game | NA |
| end\_time | Explanatory | End time of a game | NA |
| has\_game | Response | Indicates there is a game within a four-hour window | 0 = no game playing  1 = game is on |

Table-1 shows the variables details.

**Table-1 Traffic data set summary (\*\* Statistics are for the raw (as received) data set**)

There are total of 456 observations with missing values. It’s a relatively small percentage compared to our data set. To simplify the analysis, we will exclude these observations from our data set.

**Review of Literature**

There are a number of techniques that can be employed for imputation and some of them are in the list that follows:

* “Adaptive Event Detection with Time Varying Poisson Processes”, Ihler, Hutchins, Smyth, University of California, Irvine, 2006.

Selection of a method depends upon the type ort pattern associated with the missing data. Data can be missing in a completely random manner (MCAR) where there is no correlation between the missing values and those that are still present. When the data that is missing depends on the value of another variable in the data set the data is missing at random (MAR). Finally, data is determined to be missing not at random (MNAR) when the presence of data for one variable depends directly on the presence of another variable.

**Methods**

The SAS program will be used as the primary tool for performing the time series analysis and Python was uses to combine two data sets into one for analyses. Additionally, Python was used to create a dataset based on thirty minute rather than five minute intervals as well as some of the output graphics in this report.

It was discovered that the traffic monitoring device failed to record on several days. Approximately XXXX rows in the dataset were affected and it was decided to delete rather than impute the missing data. The missing rows represent only 5.9% (need to confirm) of the data set and the loss of data had not material effect on the output.

As part of the exploratory data analysis, the traffic was separated by game and non-game week days as well as game and non-game day weekends in order to determine if there were material differences in the background traffic. Detecting the presence of a baseball game requires comparison of the influence of fans leaving the stadium to the normal commuter traffic. The predictor only works if there is a difference in traffic at a time when the fans should be leaving the stadium.

With the exception of double headers, games being made up due to weather problems, and games playing late on Sundays as a featured matchup, most major league games start at 7:00 pm local time on week days and noon or 3:00 pm on weekends. The “odd” games represent a small fraction of the total games played and their inclusion simply changes the timing of the increased ramp traffic flow. In other words, regardless of then the game is played, fans leaving the stadium have the opportunity to influence the ramp traffic.

Finally, the data set has no information about venues near the stadium that may also influence ramp traffic. For this report, it is assumed that these effects do not exist during the time that ramp traffics was being captured. Traffic pattern changes that resulted from construction, accidents, weather etc. were also not considered but clearly these events could cause ramp traffic to fluctuate high or low.

**Hypothesis/Objective**

The hypothesis for this study will be to show that ramp traffic flow near a sports venue can be a predictor that an event was held.

***Ho :*** Freeway ramp traffic is a reliable predictor of an event at Dodger Stadium.

***Ha :*** Freeway ramp traffic is not a reliable predictor of an event at Dodger Stadium.

**Results**

The time series analysis included a review of ramp traffic on week days and weekend days when games were and were not being played. The graphs below illustrate the “typical” ramp traffic patterns when games were **NOT** being played. Special attention should be drawn to the time period from 6:00 pm to 11:00 pm on weekdays and 3:00 pm to 7:00 pm on weekends.



In the traffic plots where games were played, a spike in ramp traffic can be seen at approximately 11:00 pm on week nights and 8:00 pm on weekends. It takes about three hours to play a major league game and the analyses looks more closely at ramp traffic that increases one hour before the game ends and two hours afterward.

**Discussion**

**Conclusion**

Based on the results and comparisons of ramp traffic on days when games were played and not played, the conclusion is to accept the null hypothesis (**Ho**) and reject the alternative hypothesis (**Ha**). Fans leaving Dodger Stadium after baseball games, create a predictable and measurable change in ramp traffic that can be used to detect the event. Accuracy of the detection can be downgraded by accidents, weather or other events in the area and it’s possible that traffic analysis can provide a false positive.

**Future Work**

For this analysis ramp traffic was used to make a yes/no decision about whether or not an event was held at Dodger Stadium. Future analysis and insight may be gained such as predicting the size of the crowd, whether or not the Dodgers are having a winning or losing season based on crowd size, or visiting teams with featured players or strong rivalries are in town. While the traffic data was captured every five minutes, even this level of resolution is not sufficient to provide the contrasts needed to predict the aforementioned cases. Data from local businesses, in stadium concessions or other sources may be needed to help supplement a larger data set needed for this level of prediction.

**SAS Code**

**References**

SAS/STAT(R) 9.3 User's Guide. *Support.sas.com*.

Statistical Computing Seminars, Missing Data in SAS Part 1. IDRE. h*ttp://www.ats.ucla.edu/stat/sas/seminars/missing\_data/mi\_new\_1.htm*