Forecasting Crime Trends in Dallas Using Sparse Vector Auto-Regressive Modeling with Dr Shin

Team Members: John Kenney, Abdel Homi, Matt Brown, Kaushik Pasikanti

Sponsored by Dr Sunyoung Shin

TABLE OF CONTENTS

PROJECT ABSTRACT

As the title of the project implies, the original purpose of this project was to forecast crime trends in Dallas using Sparse Vector Auto Regressive (SVAR) modeling. The Dallas Police Department has historically divided portions of Dallas County into two categories, TAAG (Targeted Action Area Grids), and non-TAAG. A TAAG area, short for Targeted Action Area Grid, is a hot spot area where criminal offenses occur more frequently. TAAGs (Targeted Action Area Grids) account for 7% of the land area in Dallas County, however roughly 30% of overall crime in Dallas takes place within TAAGs. Given this difference in criminal activity, DPD has adopted different policing measures corresponding to whether a specific part of Dallas is classified as a TAAG area or non-TAAG area.

The purpose of the project was to forecast crime offenses (I.e., assault, burglary, theft, etc.) in both areas over time and compare the results. The team examined various methods including SVAR modeling (the original approach), traditional VAR modeling, and AR modeling (modeling by each crime type independently). Both SVAR and VAR models relate present observations of a particular variable (crime type) with past observations of that crime type and past observations of other crime types in the dataset. Incorporating each of these approaches, instead of exclusively the original SVAR model helped us gain a more comprehensive understanding of the dataset. The SVAR model used on the entire dataset produced the most accurate predictions using the weighted mean forecast error as our error metric.

INTRODUCTION

The purpose of this project from a high-level overview was to forecast crime trends in TAAG and non-TAAG areas present in Dallas County, using the sparse vector auto regression forecasting technique. For this project, we are primarily concerned about comparing changes in criminal activity by crime type in TAAG areas where crime is more concentrated, relative to non-TAAG areas where crime is less frequent. Traditional VAR models are unable to describe relationships between different predictors, hence the need for a sparse VAR model. In our context, a sparse VAR model will be able to account for the potential relationships between various crime types. For example, when trying to forecast the future value of homicides for a particular TAAG area, it may be helpful to be aware that homicides are strongly correlated to drug offenses. An SVAR model will take this kind of hypothetical example into account, while a VAR model would not.

Forecasting crime trends with SVAR modeling in TAAG and non-TAAG areas within Dallas County is a specific example in the broad category known as time-series prediction. A time-series prediction problem uses historical time-stamped observations to predict future observations. Popular applications of time-series prediction include the use of stock market data to predict the price movement of specific security, meeting staffing requirements for a call center based on forecasted demand, etc.

Other solutions to a time-series prediction problem include the use of machine learning techniques. The existence of deep learning methods offers a lot of promise for time series forecasting problems. These methods automatically learn of temporal dependence, and they automatically handle temporal structures like trends and seasonality. In our case, we had to write code to account for these realities in the original dataset. Abdel and Kaushik discussed the merits of considering at approach such as the LSTM method, which remembers its previous cell states that can have great predictive power when dealing with data such as stock market data.

COMMUNICATION PLAN

Our team met with Dr. Shin every Tuesday afternoon for an hour and fifteen minutes. During the first half of the semester, we met at 11:30 am, and due to a change in work schedules, we met at 1:30 pm through November and December up to the conclusion of the project. We met as a team on Saturday afternoons to review the progress of the project and collaborate. Any intermittent meetings took place on Discord.

Throughout the semester, our team communicated via phone/text, and we communicated any questions or concerns about the project to Dr. Shin via email. The team members communicated using the popular instant messaging and voice calling platform called Discord throughout the semester. Using Discord, we shared updates with how various parts of the project were progressing, and we used it to meet virtually when meeting in person was not feasible. This is a free to use platform, and it is extremely helpful when collaborating with other students.

Communication within the group was difficult at first as team members had different ideas about the best direction to proceed in working on the project (I.e., which programming environment to work in). With the assistance of the professor, we were able to establish regular meeting times outside of the meetings with Dr. Shin to collaborate on the project. We ended up meeting on Saturday afternoons at a coffee shop in downtown Plano.

TIMETABLE

(Milestones, deliverables clearly defined)

- Exploratory data analysis and data cleaning in R completed by late September
- SVAR, VAR, and AR models were completed by early November.
- Forecasting and error measures for each of the above models were implemented in the middle of November.
- Model evaluation completed by end of November.
- Final
- Mapping visualization (not heatmap) of TAAG areas in Dallas was completed by early November.

IMPLEMENTATION DETAILS

Abdel and Kaushik worked with the crime data in a Python notebook environment, and they were primarily responsible for generating visualizations of the city of Dallas, and crime data related to the different areas, TAAG and non-TAAG in Dallas. For example, Abdel and Kaushik used a Python library called GeoJSON to implement a map of the TAAG areas present in the city of Dallas. This mapping visualization was complemented with a heatmap to better perceive the increase in crime counts per TAAG area year over year.

Matt and John worked in RStudio to clean the data and transform the data to make it suitable for modeling. During this semester, we had to switch raw data files after realizing that our original raw data file had a missing data gap for crime data that accounted for three years of data. After identifying this gap present in the original raw data file, we selected a new raw data file and confirmed we did not have the same issue.

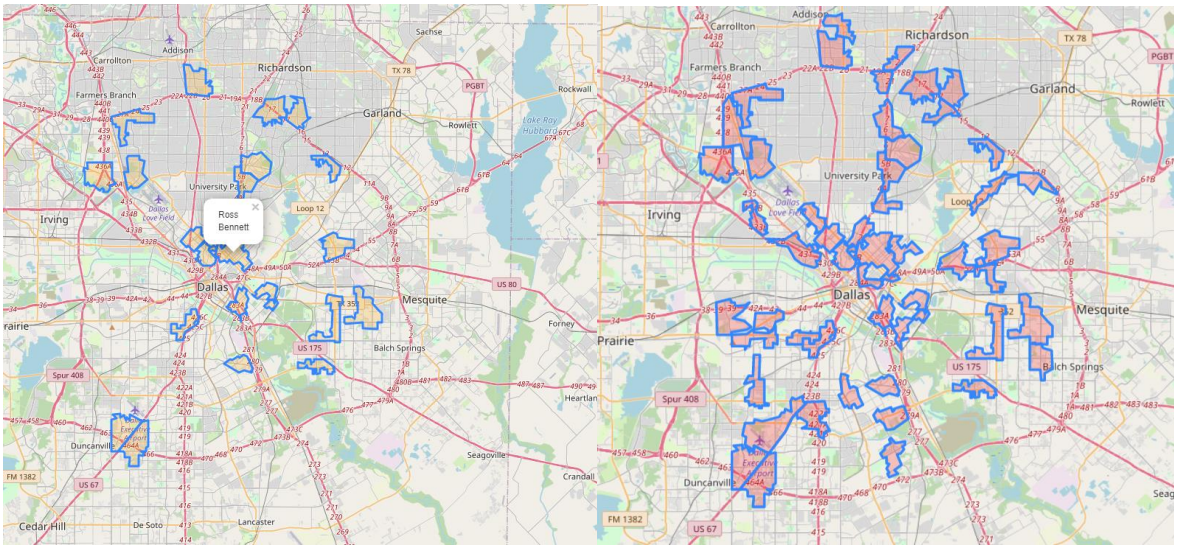
We first had to clean our data from the raw file. To do this we grouped each incident into either ASST (Assault), BURG (Burglary), DRUG (Drug), HOMD (Homicide), ROBB (Robbery), THEF (Theft), WCCR (White Collar Crime), and Others. We dropped Others because the incidents in this category were not easily explained. We considered modeling our data on monthly data and weekly data by aggregating the crime counts of each category. We found that monthly data when modeling did not have an adequate number of observations to find meaningful relationships in our data. We further created two more data sets for both monthly and weekly data that split the overall crime counts into TAAG and Non-TAAG data sets.

The Dallas Police Department lost an estimate of 1 trillion terabytes of data in 2021. Therefore, we only included the years 2014 through 2020. Moreover, we have faced an issue in extracting the right and updated taag areas. As I was creating the maps I have download the latest taag areas from DPD with a valid DISCLAIMER. Interestingly, we have found a major difference between both maps. After, we implemented an algorithm that compares both datasets. We only found 22 matched taag areas. Hence, we went ahead and performed the analysis on only the updated areas.

The figures below are confirmation of the difficulties we have faced we the provided and updated datasets. The first graph represents the di represents the difference between initial taag dataset vs. Updated dataset.

```
print('capstone_initial_taags: ', len(initial_taags)), print('json_updated_taag: ', len(updated_taag))
```

```
capstone_initial_taags:  31
json_updated_taag:  54
```



```
taags_union  = list(set(initial_taags) & set(updated_taag))
```

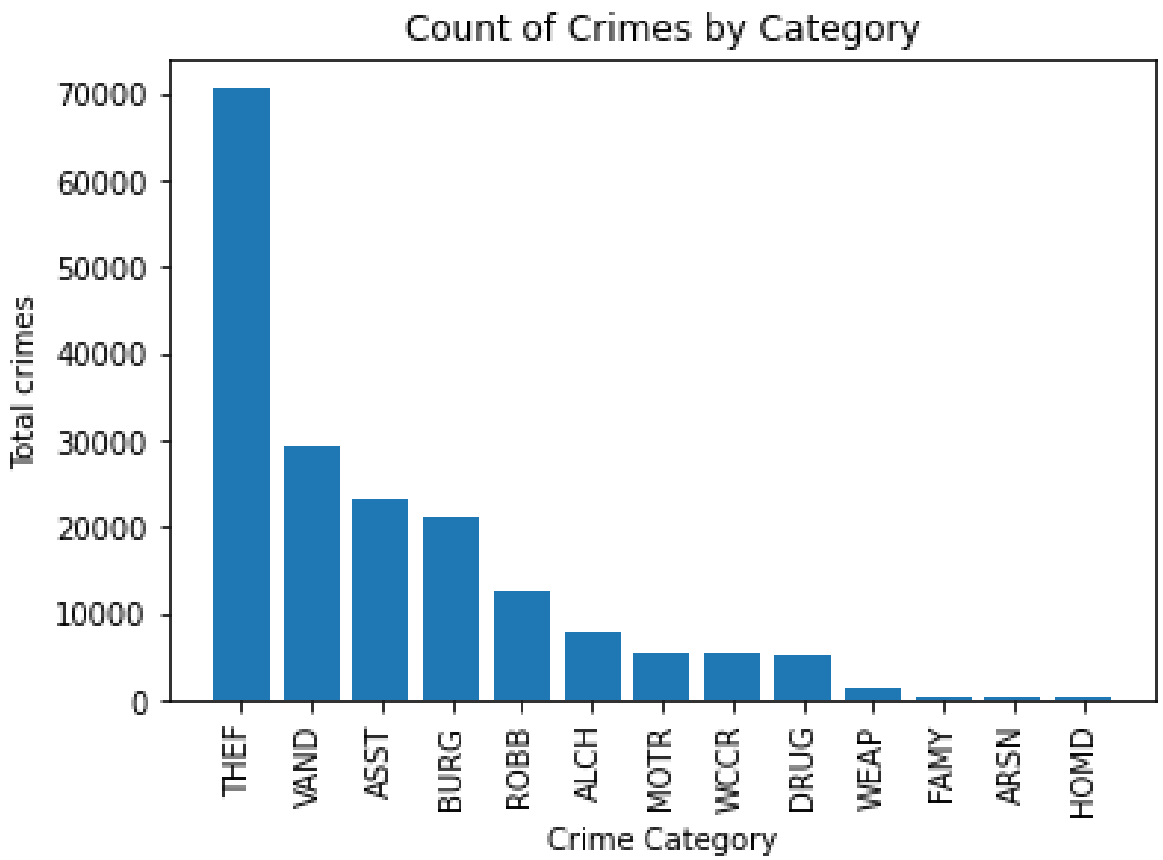
```
taags_union
```

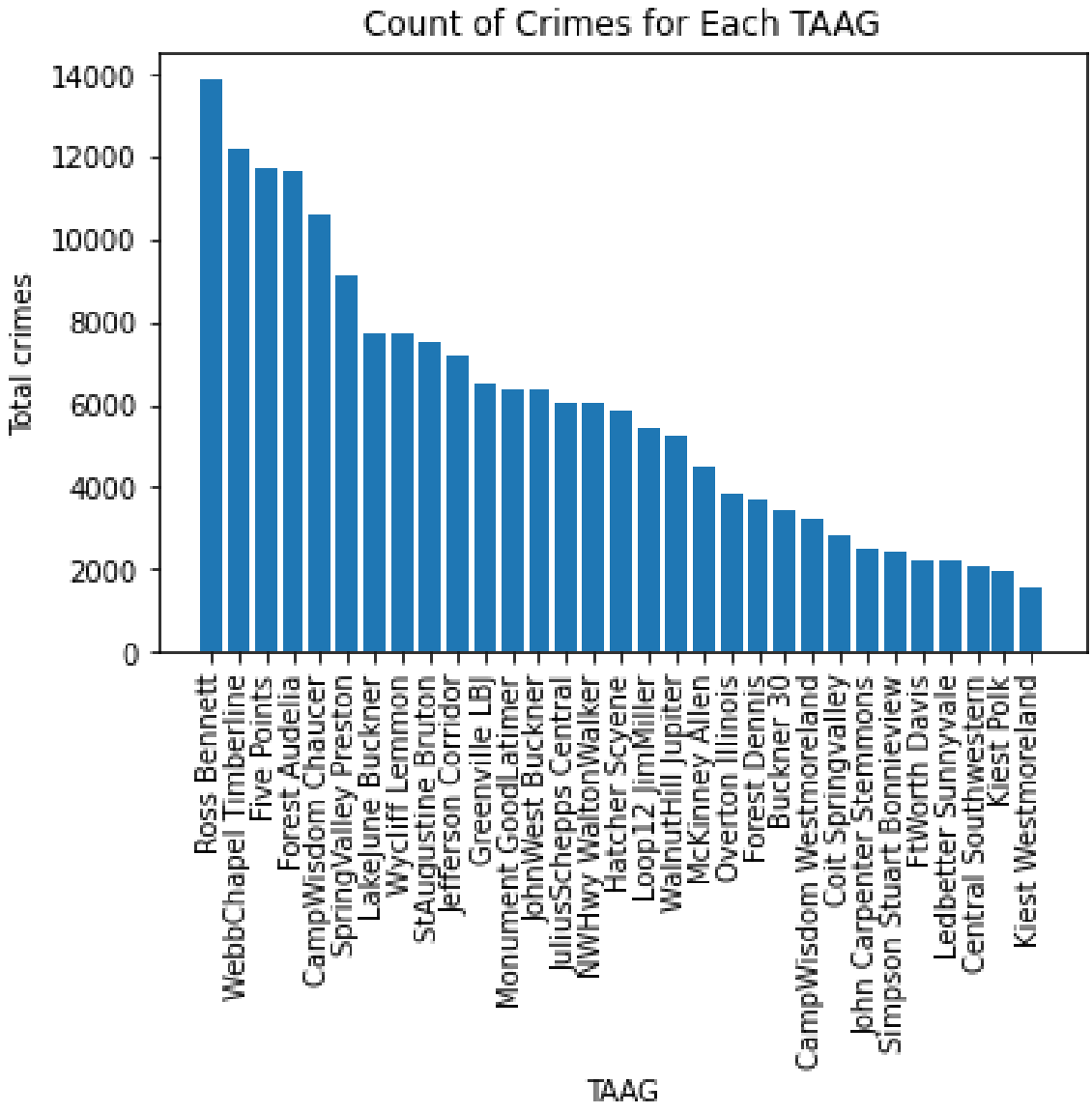
```
['Forest Dennis',
 'Forest Audelia',
 'Ross Bennett',
 'SpringValley Preston',
 'WalnutHill Jupiter',
 'Greenville LBJ',
 'JohnWest Buckner',
 'WebbChapel Timberline',
 'Loop12 JimMiller',
 'Central Southwestern',
 'CampWisdom Chaucer',
 'Hatcher Scyene',
 'McKinney Allen',
 'Five Points',
 'StAugustine Bruton',
 'Wycliff Lemmon',
 'NWHwy WaltonWalker',
 'CampWisdom Westmoreland',
 'Jefferson Corridor',
 'Overton Illinois',
 'JuliusSchepps Central',
 'LakeJune Buckner']
```

Visualization:

The following graph is a clear distinction between all types of crimes within all taag areas. From this plot we can draw an attention to newly added plots such as Mckiney/Allen, Forest Dennis, Capswisdom Westmoreland where they were added and considered later within the period we analyzed and modeled.

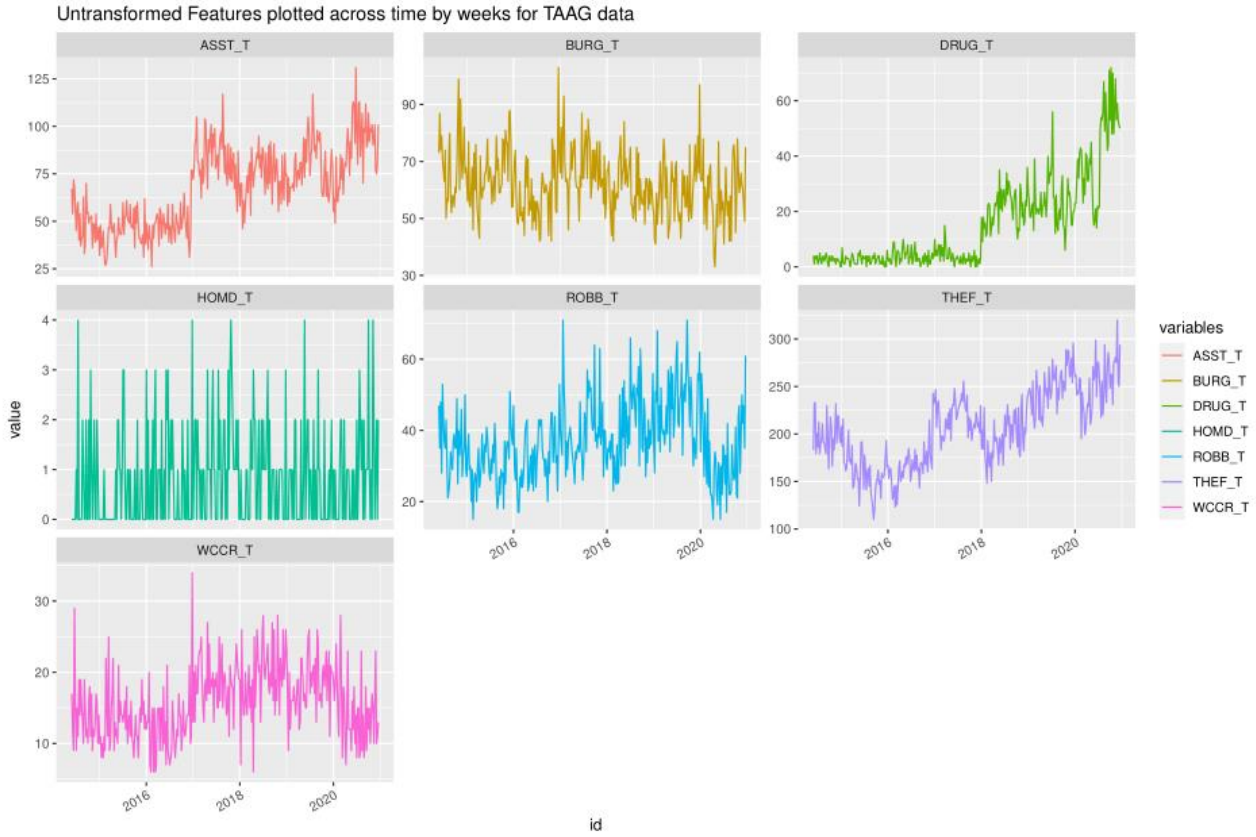

TAAG = Jefferson Corridor
TAAG = SpringValley Preston
TAAG = Greenville LBJ
TAAG = Ross Bennett
TAAG = Hatcher Scyene
TAAG = NWHwy WaltonWalker
TAAG = McKinney Allen
TAAG = JohnWest Buckner
TAAG = Loop12 JimMiller
TAAG = LakeJune Buckner
TAAG = Five Points
TAAG = WalnutHill Jupiter
TAAG = Wycliff Lemmon
TAAG = CampWisdom Chaucer
TAAG = Overton Illinois
TAAG = StAugustine Bruton
TAAG = WebbChapel Timberline
TAAG = Forest Dennis
TAAG = Central Southwestern
TAAG = JuliusSchepps Central
TAAG = CampWisdom Westmoreland
TAAG = Forest Audelia
count
Date
NIBRS_Crime_Category
WEAP
HOMD
ARSN
FAMY
MOTR
VAND
DRUG
WCCR
BURG
ALCH
ASST
THEF
ROBB

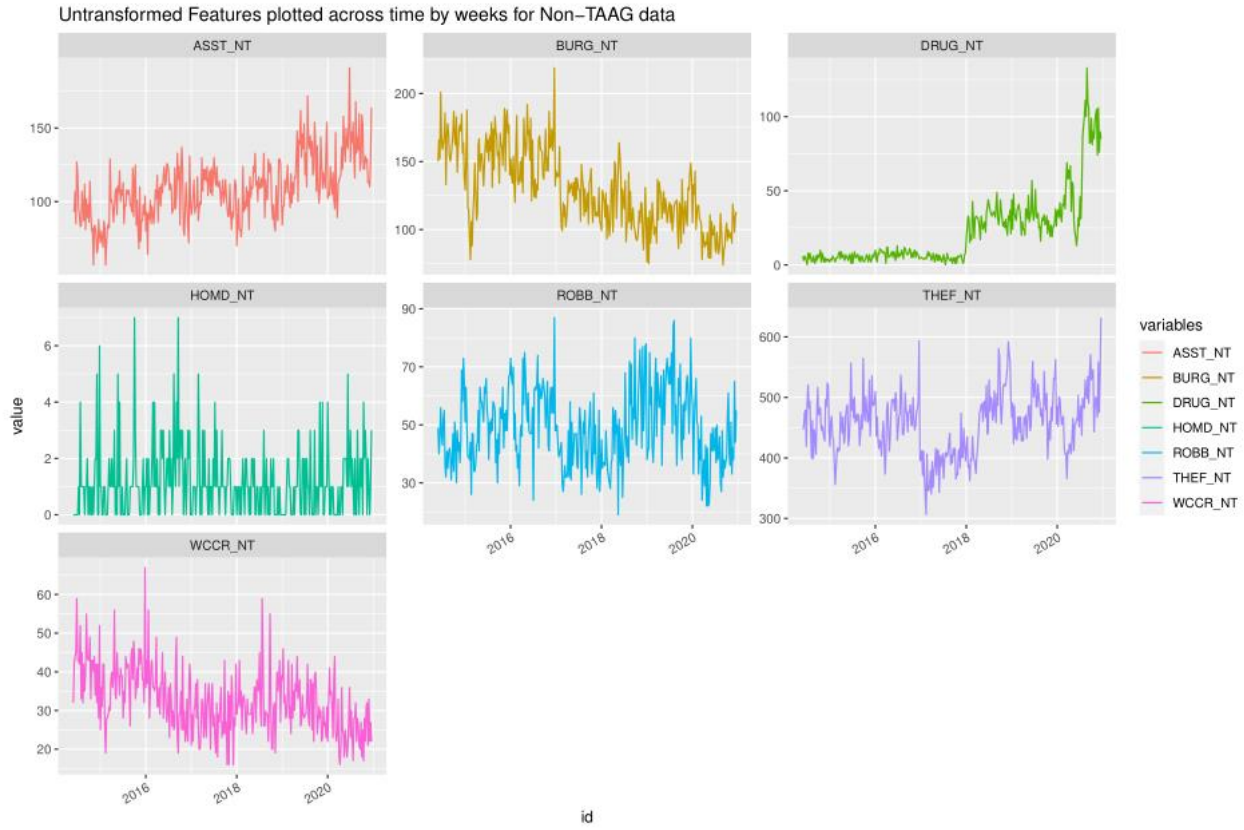

Count of Crimes by Category
Total crimes
70000
60000
50000
40000
30000
20000
10000
0
THEF
VAND
ASST
BURG
ROBB
ALCH
MOTR
WCCR
DRUG
WEAP
FAMY
ARSN
HOMD
Crime Category

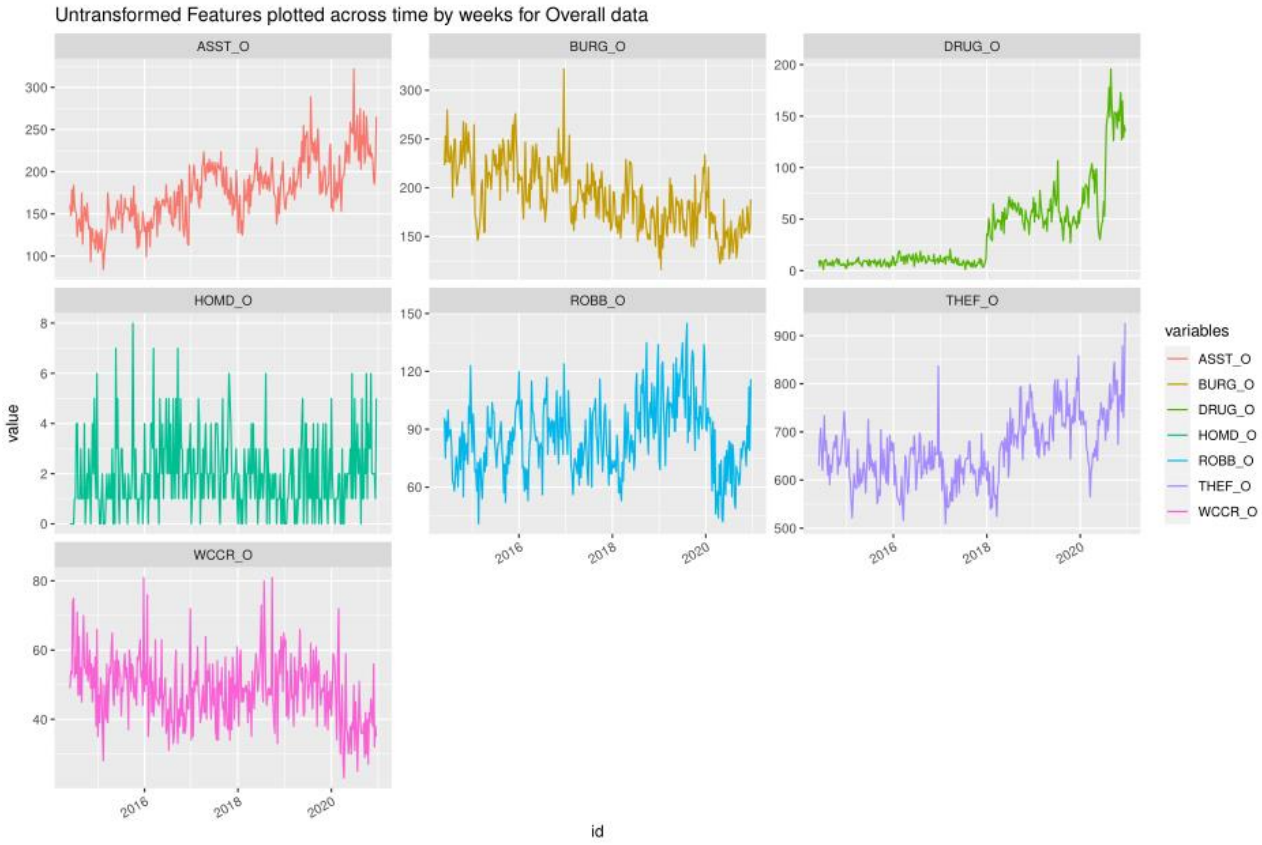


The chart above lets us see which TAAG areas are historical hot spots.

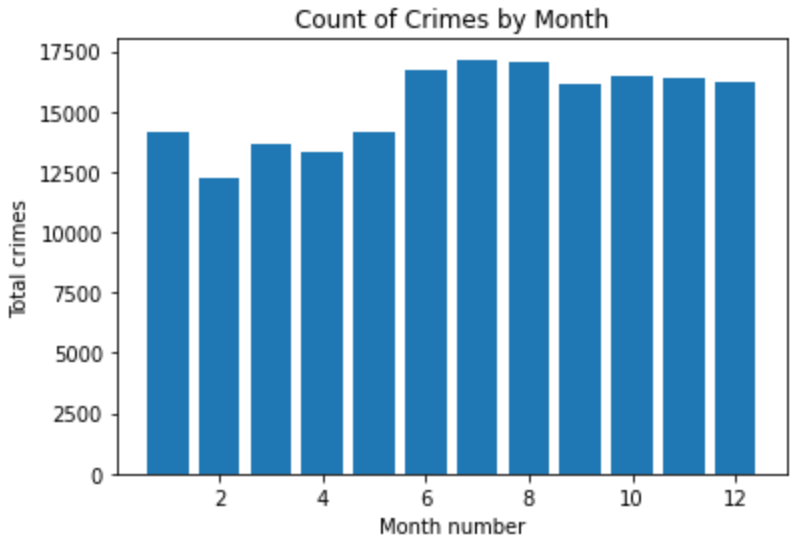
Graph below plots each time series


Untransformed Features plotted across time by weeks for TAAG data
ASST_T
BURG_T
DRUG_T
HOMD_T
ROBB_T
THEF_T
WCCR_T
value
id
2016
2018
2020
variables
ASST_T
BURG_T
DRUG_T
HOMD_T
ROBB_T
THEF_T
WCCR_T

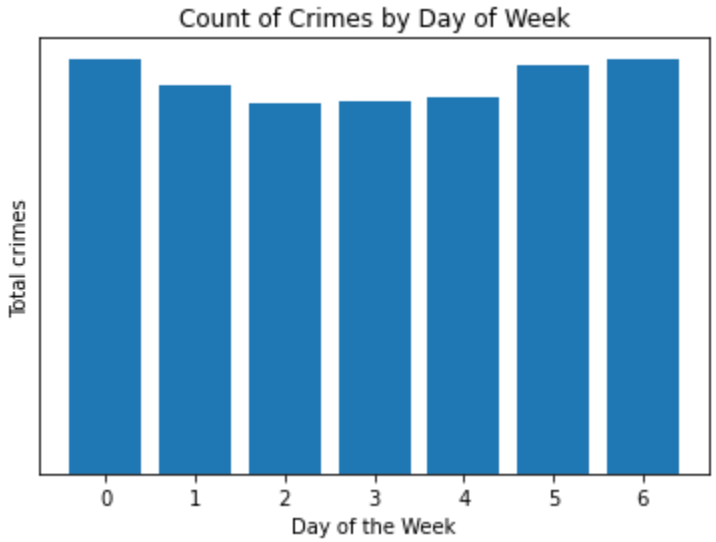

Untransformed Features plotted across time by weeks for Non−TAAG data
ASST_NT
BURG_NT
DRUG_NT
HOMD_NT
ROBB_NT
THEF_NT
WCCR_NT
value
id
variables
ASST_NT
BURG_NT
DRUG_NT
HOMD_NT
ROBB_NT
THEF_NT
WCCR_NT
2016
2018
2020



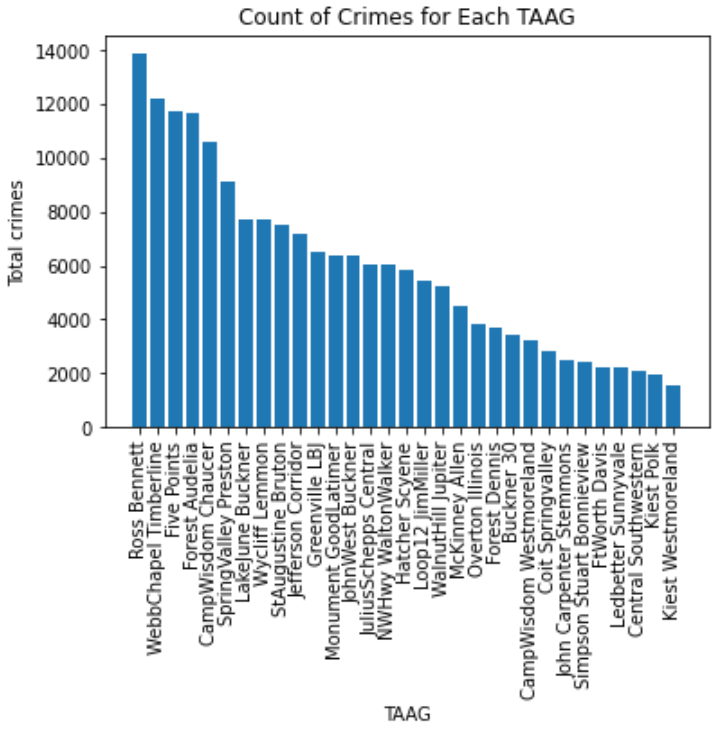
We plotted general plots of the crime count by certain date metrics so we can visualize seasonality patterns. This helps us in picking what models would be of best fit as well as visualize and patterns in the data that we would otherwise miss.



Plotting the crime count by month over the years, lets us visualize seasonality patterns in the data, which may help with modeling aspects. We can see that more crimes tend to happen during the summer months. There are many reasons that this may be true. For example, because the days are longer, more people are outdoors and, on the streets, later at night, leaving more opportunities for crimes. Another reason may be because teens are out of school, and many are unsupervised so one can expect an increase in gang activity and juvenile crimes. A Lot more social events, where alcohol consumption is prevalent, tend to occur during the summer months over the spring and winter months.

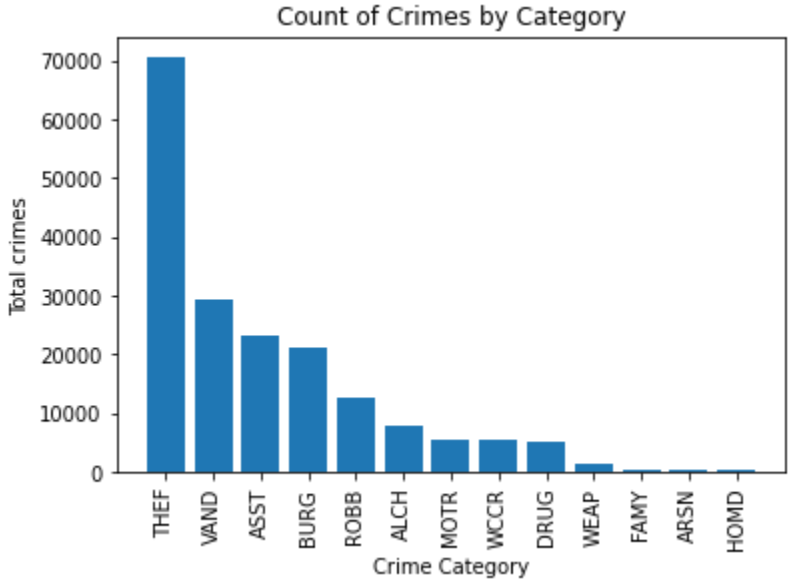


I had some technical difficulties adding the x-label ticks to label the numbers as weekdays instead, but 0 corresponds to Sunday and Friday corresponds to Saturday. From The plot above we can see that more crimes tend to happen on the weekends (Friday, Saturday, and Sunday). This makes sense as more people are out on the streets on Friday and Saturday nights, which may lead to Sunday early morning. More people go out to bars, clubs, and other social events where they may get intoxicated, leading to arguments, then fights, followed by stabbings and shootings.



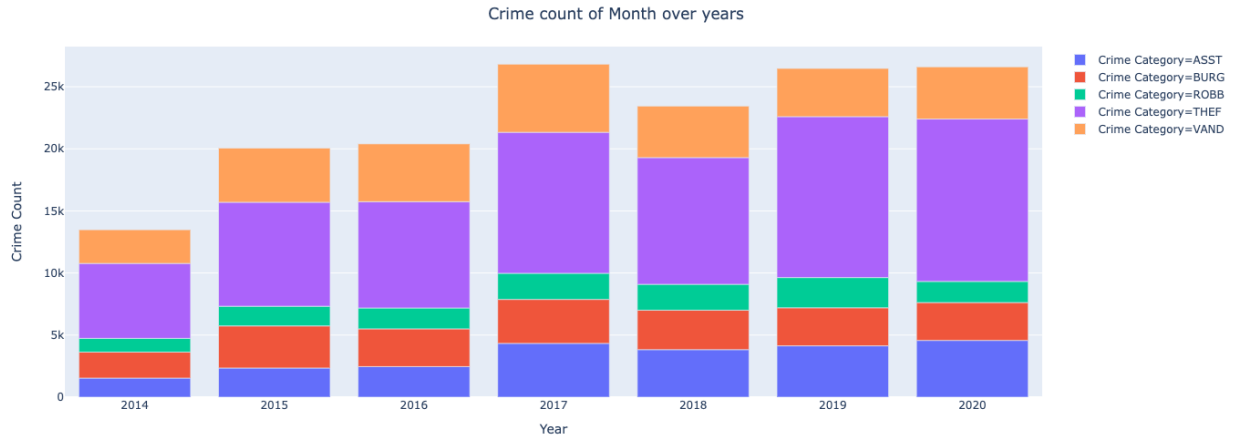
This figure explains the spread of crimes in each TAAG area in Dallas from 2014-2020.

This shows which TAAG's have been historical crime hotspots compared to TAAG areas that may have been discovered more recently as hotspots by that Dallas Police Department.

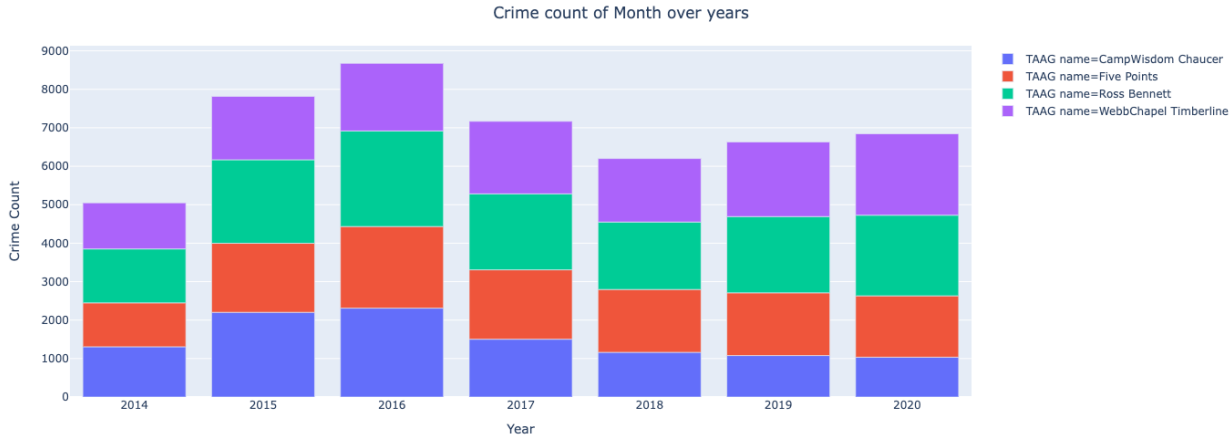


This figure shows us the crime categories that we were given, and which type of crimes occur the most.

Crimes against property, theft, and burglary are especially prevalent in downtown areas of metro cities, so this makes sense to see. These types of crimes tend to happen a lot in commercialized areas, or areas with a lot of shops, offices, theatres, and restaurants, which downtown Dallas has a high concentration of.



Here we are analyzing the trend of the counts of the top 5 crime categories over the years. I was surprised to see that the total crime count did not change in 2019 and 2020 as those two years were affected heavily by covid. I initially assumed that because less people were going out, the count of crimes such as Assault, Robbery, Theft would, be a lot less, but that did not seem to affect the crime counts of these types of crimes as the total crime count of 2019 and 2020 respectively, are greater than the total crime counts in 2015, 2016, and 2018. The proportion of each crime category counted to the total crime count stayed the same.



Here we are plotting the top 4 TAAGS with the highest total crime count and plotting the total crime count for each per year with stacked bar plot. To enhance this graph, we should include the percentage out of the total crime count.

After cleaning our data, we then had to transform each of our data sets to be stationary. The reason for doing this is so we can eliminate trends in our data. The transformations applied to data were as follows, we first attempted a log transformation if there were no zeros in the timeseries and then whether we were able to log or not we took the difference

of each observation. We further confirmed that our data was stationary by performing a Phillips-Perron Test for Unit Roots. An example on the TAAG dataset can be seen below in Table 1 and 2.

Table 1: stationary or not?

	Dickey-Fuller	Truncation lag parameter	p-value
ASST_T_diff(log)	-39.75676	5	0.01
BURG_T_diff(log)	-36.69492	5	0.01
DRUG_T_diff()	-33.68827	5	0.01
HOMD_T_diff()	-49.28785	5	0.01
ROBB_T_diff(log)	-39.79157	5	0.01
THEF_T_diff(log)	-35.24954	5	0.01
WCCR_T_diff(log)	-54.21746	5	0.01

Note:
Test used Phillips-Perron Test for Unit Roots
[1] P-value < 0.05 means data is stationary
[2] P-value > 0.05 means data is not stationary so try a different method.

Table 2: Table of Features and the Transformations Applied on Weekly TAAG data

Variable_W	Description_W	Transformation_Applied_W
ASST_T	Assault related Crimes in all of DALLAS	diff(log)
BURG_T	Burglary related Crimes in all of DALLAS	diff(log)
DRUG_T	Drug related Crimes in all of DALLAS	diff()
HOMD_T	Homicide related Crimes in all of DALLAS	diff()
ROBB_T	Robbery related Crimes in all of DALLAS	diff(log)
THEF_T	Theft related Crimes in all of DALLAS	diff(log)
WCCR_T	White Collar related Crimes in all of DALLAS	diff(log)

Var analysis:

Next, we fitted the Sparse Var model w/ HLag penalty using the full data. We found that that the TAAG dataset found relationships between other crimes, but the non-TAAG only formed an AR (1) model finding that only each timeseries' previous step had an influence on each crime. Also, we found that while the TAAG model found relationships between other crimes that based on its coefficients each crime was strongly related to its previous observation. We can further explore the relationships between our models for TAAG and non-TAAG by examining the coefficients from tables 7-13 for TAAG and tables 14-20 for non-TAAG.

Table 7: TAAG model response variable is Assault

Variables	Coefficients
ASST_T: Lag 1	-0.4675
ASST_T: Lag 2	-0.1433
Intercept	0.0063

Table 8: TAAG model response variable is Burglary

Variables	Coefficients
BURG_T: Lag 1	-0.3957
WCCR_T: Lag 1	-0.0237
BURG_T: Lag 2	-0.1064
BURG_T: Lag 3	-0.0138
Intercept	-0.0022

Table 9: TAAG model response variable is Drug

Variables	Coefficients
DRUG_T: Lag 1	-0.3786
THEF_T: Lag 1	-0.0107
THEF_T: Lag 2	0.0032
THEF_T: Lag 3	-0.0001
Intercept	0.0031

Table 10: TAAG model response variable is Homicide

Variables	Coefficients
HOMD_T: Lag 1	-0.4858
WCCR_T: Lag 1	0.0398
HOMD_T: Lag 2	-0.1741
WCCR_T: Lag 2	-0.0253
HOMD_T: Lag 3	-0.0185
HOMD_T: Lag 4	-0.0037
Intercept	-0.0026

Table 11: TAAG model response variable is Robbery

Variables	Coefficients
BURG_T: Lag 1	-0.0090
ROBB_T: Lag 1	-0.4813
THEF_T: Lag 1	0.0068
BURG_T: Lag 2	0.0015
ROBB_T: Lag 2	-0.1624
ROBB_T: Lag 3	-0.0376
Intercept	0.0035

Table 14: Non-TAAG model response variable is Assault

Variables	Coefficients
ASST_NT: Lag 1	-0.2561
Intercept	0.0070

Table 15: Non-TAAG model response variable is Burglary

Variables	Coefficients
BURG_NT: Lag 1	-0.2575
Intercept	-0.0055

Table 16: Non-TAAG model response variable is Drug

Variables	Coefficients
DRUG_NT: Lag 1	-0.2339
Intercept	0.0045

Table 17: Non-TAAG model response variable is Homicide

Variables	Coefficients
HOMD_NT: Lag 1	-0.3397
Intercept	-0.0115

Table 18: Non-TAAG model response variable is Robbery

Variables	Coefficients
ROBB_NT: Lag 1	-0.320
Intercept	-0.004

Table 12: TAAG model response variable is Theft

Variables	Coefficients
BURG_T: Lag 1	-0.0074
THEF_T: Lag 1	-0.3749
BURG_T: Lag 2	0.0016
THEF_T: Lag 2	-0.1162
THEF_T: Lag 3	-0.0127
Intercept	-0.0068

Table 19: Non-TAAG model response variable is Theft

Variables	Coefficients
THEF_NT: Lag 1	-0.1696
Intercept	-0.0035

Table 13: TAAG model response variable is White Collar Crime

Variables	Coefficients
DRUG_T: Lag 1	-0.0121
ROBB_T: Lag 1	0.0167
WCCR_T: Lag 1	-0.5782
WCCR_T: Lag 2	-0.2086
WCCR_T: Lag 3	-0.0397
Intercept	0.0007

Table 20: Non-TAAG model response variable is White Collar Crime

Variables	Coefficients
WCCR_NT: Lag 1	-0.3026
Intercept	-0.0006

Model Evaluation:

The last step we took as part of our project was model evaluation. To do this we considered 3 models: sparse var w/ HLag penalty, regular var, and an AR (1) model. Under the direction of Professor Shin we used the method of non-rolling out of sample model evaluation. To perform this evaluation, we needed to be able to forecast 1 step ahead with new data. This required us to right our own forecast function for both the sparse var and regular var model. The math formulation of the operation can be seen below.

out of sample forecast 1 step ahead equation

$$\begin{pmatrix} \hat{Y}_1(T+1) \\ \hat{Y}_2(T+1) \\ \vdots \\ \hat{Y}_k(T+1) \end{pmatrix} = \Phi_{k,L\times k} \cdot \begin{pmatrix} Y_1(T) \\ Y_2(T) \\ \vdots \\ Y_k(T) \\ \vdots \\ Y_1(T-L+1) \\ Y_2(T-L+1) \\ \vdots \\ Y_k(T-L+1) \end{pmatrix} + \Phi_{0_{k,1}}$$

We also used the error metric shown below to evaluate our models.

RMSE

$$RMSE = \frac{1}{k \times (T_2 - T_1)} \sum_{i=1}^{k} \sum_{t=T_1}^{T_2-1} \sqrt{(\hat{y}_{i,t+1} - y_{i,t+1})^2}$$

And further breakdown of our model evaluation can be seen using the following metric.

$RMSE_i$

$$\begin{pmatrix} RMSE_1 = \frac{1}{(T_2-T_1)} \sum_{t=T_1}^{T_2-1} \sqrt{(\hat{y}_{i,t+1} - y_{i,t+1})^2} \\ \vdots \\ RMSE_k = \frac{1}{(T_2-T_1)} \sum_{t=T_1}^{T_2-1} \sqrt{(\hat{y}_{k,t+1} - y_{k,t+1})^2} \end{pmatrix}^T$$

After formulating our forecast method and error metric we then split our data using the first 80% of our observations as training data and the remaining 20% as test data. To scale our data before doing the evaluations we first scaled the training data and then used the same attributes to scale the test data to not get a biased evaluation. We also had to convert all our forecasted values back to original count values to do this we unscaled and then applied the inverse of the transformations applied to get the data to stationary. The reason we did this was to be able to break down our forecasts from the overall model to TAAG and non-TAAG crime counts. To do this we calculated the prior probability of TAAG crimes and non-TAAG crimes in our training data. We considered the straightforward way to calculate the prior probabilities for each crime by summing the count in training TAAG data and dividing by the training Overall data. To get the prior probabilities for non-TAAG crimes for each probability we did 1 - prior probability of TAAG crime. E.g prior(TAAG_assualt) = sum(TAAG_assualt) / sum(Overall_assualt), prior(nonTAAG_assualt) = 1 – prior(TAAG_assualt). Once we got the prior probabilities, we multiplied the overall forecasted values by each category's prior probability to get forecasted TAAG and non-TAAG values from the overall model. When comparing each methods forecasted values after unscaling and then untransforming we actually had to subtract the observed value with an additional + 1, because since we undifferenced the data we gained an observation of xi to our data. This only talking about how it is coded in the code because we had to adjust the code to have the error formula work since we are adding an observation to the beginning of our dataset. With this we were able to calculate the overall model's performance for forecasting crimes in TAAG and non-TAAG areas. The model evaluations showed that the sparse var w/ HLag penalty performed the best. As can be seen in table 28. Further tables and graphs can be seen in the codeall.pdf which includes tables and graphs for all three datasets.

Table 28: Out-Of-Sample model evaluation

	mean(RMSE on TAAG DATA)	mean(RMSE on NON-TAAG DATA)
SparseVar_CV_HLAG	15.68545	25.99907
SparseVar_CV_HLAG_Overall	19.26385	28.05909
VAR_AIC	18.21052	28.76986
VAR_AIC_Overall	19.80248	29.02669
AR(1)	17.06835	26.91332
AR(1)_Overall	19.53949	28.64023

Note:
Train/Test split ratio 0.8
Note:
Prior probabilities of TAAG/Overall for each crime to get the forecast of crimes
in TAAG areas from Overall model forecasted values.
It follows that the prior probability for Non-TAAG is 1 - the priors for TAAG/Overall
[1] TAAG Prior for Assault: 0.38
[2] TAAG Prior for Burglary: 0.32
[3] TAAG Prior for Drug: 0.4
[4] TAAG Prior for Homicide: 0.41
[5] TAAG Prior for Robbery: 0.43
[6] TAAG Prior for Theft: 0.3
[7] TAAG Prior for White Collar Crime: 0.32

For further analysis we have the RMSEi's for TAAG and non-TAAG that let us see which model worked best for each crime. This can be seen in tables 29 and 30.

Table 29: Out-Of-Sample model evaluation RMSE for TAAG data for each crime

	ASST_T	BURG_T	DRUG_T	HOMD_T	ROBB_T	THEF_T	WCCR_T
SparseVar_CV_HLAG	17.10322	11.69895	28.56807	1.322444	19.08573	26.15931	5.860448
SparseVar_CV_HLAG_Overall	21.92370	15.40216	30.51437	1.165267	14.22669	44.41859	7.196149
VAR_AIC	20.83907	13.79238	29.96194	1.629114	22.44529	32.16905	6.636778
VAR_AIC_Overall	22.20559	16.03347	30.68316	1.181782	14.83752	45.70529	7.970517
AR(1)	18.83788	13.16226	29.13872	1.492289	21.70193	28.41398	6.731422
AR(1)_Overall	22.24183	15.79020	30.49776	1.204358	14.42627	45.02993	7.586116

Note:
Train/Test split ratio 0.8
Note:
Prior probabilities of TAAG/Overall for each crime to get the forecast of crimes
in TAAG areas from Overall model forecasted values.
[1] TAAG Prior for Assault: 0.38
[2] TAAG Prior for Burglary: 0.32
[3] TAAG Prior for Drug: 0.4
[4] TAAG Prior for Homicide: 0.41
[5] TAAG Prior for Robbery: 0.43
[6] TAAG Prior for Theft: 0.3
[7] TAAG Prior for White Collar Crime: 0.32

Table 30: Out-Of-Sample model evaluation RMSE for Non-TAAG data for each crime

	ASST_NT	BURG_NT	DRUG_NT	HOMD_NT	ROBB_NT	THEF_NT	WCCR_NT
SparseVar_CV_HLag	22.62760	24.78445	50.08702	1.439145	18.90340	50.19170	13.96014
SparseVar_CV_HLAG_Overall	25.00690	21.05555	47.65669	1.369274	26.73123	64.38042	10.21358
VAR_AIC	24.89668	26.40771	52.46025	1.667385	22.90393	57.24167	15.81140
VAR_AIC_Overall	26.77407	21.15895	47.92450	1.502647	29.06002	64.78078	11.98588
AR(1)	24.43368	25.77211	50.29544	1.535721	20.10323	51.52956	14.72351
AR(1)_Overall	26.52192	21.73032	47.63698	1.424977	27.71585	64.60533	10.84627

Note:

Train/Test split ratio 0.8

Note:

Prior probabilities of Non-TAAG/Overall for each crime to get the forecast of crimes in Non-TAAG areas from Overall model forecasted values.

[1] Non-TAAG Prior for Assault: 0.62

[2] Non-TAAG Prior for Burglary: 0.68

[3] Non-TAAG Prior for Drug: 0.6

[4] Non-TAAG Prior for Homicide: 0.59

[5] Non-TAAG Prior for Robbery: 0.57

[6] Non-TAAG Prior for Theft: 0.7

[7] Non-TAAG Prior for White Collar Crime: 0.68

ISSUES AND LESSONS LEARNED

One of the issues that we struggled with during the first half of the semester was communication around meeting together each week (i.e. when, where, how often, etc.). This was resolved after a brief meeting with our professor, and we found that meeting Saturday afternoons at a coffee shop was a good fit. John and Matt were in regular communication throughout each week as the R coding portion of the project progressed, while Abdel and Kaushik communicated with each other for the Python programming portion of the project. Questions around the direction of the project and adapting the project to include a wider variety of modeling techniques were clarified by Dr. Shin as our goals evolved throughout the semester.

ABDEL: Also, another issue we ran into was the difference in the number of TAAG areas depending on the data file used. We resolved this issue by only using the TAAG areas that were present in both our initial data file and our updated data file which was used to produce the initial mapping visualizations. This was an issue specific to the Python portion of the project where we visualized the TAAG areas across a Google Maps view of Dallas County.

We identified that the original dataset was corrupted by a significant loss of data that was accidentally deleted by the DPD.

The first issue faced was in modeling the data using monthly data we found that all our coefficients came out as zero when using the sparse var model. The reason this was the case was we had little observations to find any meaningful relationships in our data. From this I learned the importance of having a larger data to model with.

John: Lastly, the hardest thing I faced was the model evaluation. Having to transform the forecasts back to original counts was a complicated challenge on top of forecasting values using new data. The reason we had to scale our data was that sparsevar model did not work well with unscaled data. It often times only gave zeros as coefficients if the data was not scaled. Therefore, we needed to scale our data, but we had to scale the testing data using the scaled attributes calculated from scaling the training data. This way we are not corrupting the test set for model evaluation. We also needed to go back to crime count numbers so we could evaluate how well our overall model did in comparison to the models evaluated on only taag and non-taag data. The reason being is that we needed to break down the overall forecasted values to taag and non-taag forecasts that would give us errors comparable to our other models. Model evaluation became a little trickier because by

undifferencing our data we had another observation in our data we had to account for in the original observations. It took me some time to account for this problem, but I finally figured out that I had to index an additional step ahead in the observed test set before calculating the RMSE, because we are adding a value to the beginning of the dataset. I learned how to get values back to the original form and perform model evaluation on them.

John: I learned the steps involved when working with time series data, and how to use models such as sparse var, var, and AR (1).

CONCLUSION

(Why is this a good project, what is the usability? Is there any?)

From a conceptual perspective, being able to better pinpoint the locations of more frequent criminal activity occurrences results in more effective use of policing resources in the areas of Dallas where this activity takes place. Looking at the project, what we have done up to this point would need to be expanded prior to being used by the DPD and implementing different policing practices in certain areas of the city. By better predicting when an uptick of a certain crime will occur will enable DPD to stop crimes as they occur or stop them before they happen.

EVALUATION

As detailed in the TAAG.pptx provided by Dr. Shin at the beginning of the semester, the original goals of the project were as follows:

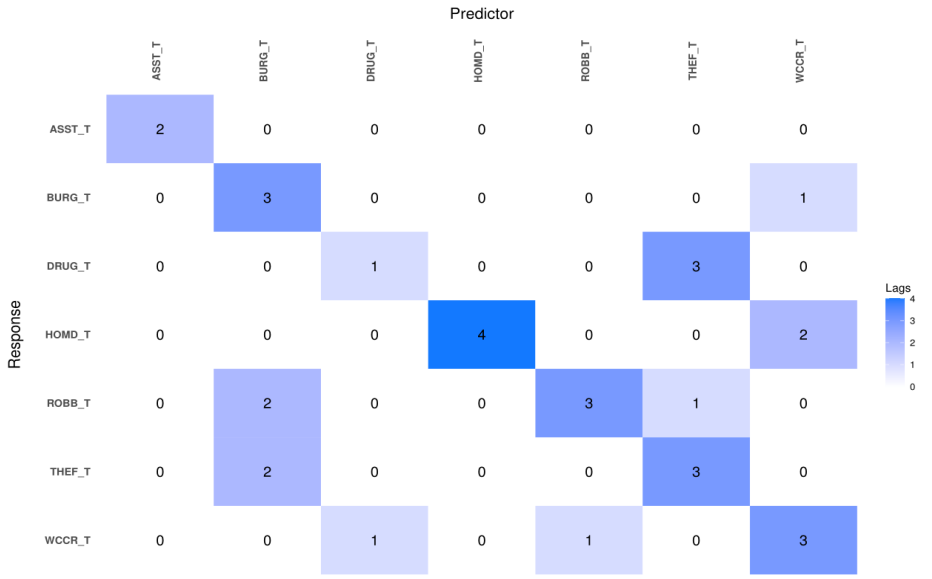
- Transform the data to make it stationary.
- Fit a sparse VAR model to the data using the R package bigtime
- Draw a heatmap for the results of the SVAR.

The project was successful in that we were able to implement the three original goals mentioned above and outlined in the original PowerPoint provided by Dr. Shin. Each of the original goals were accomplished in R, and additional goals were added to the project that were completed in Python.

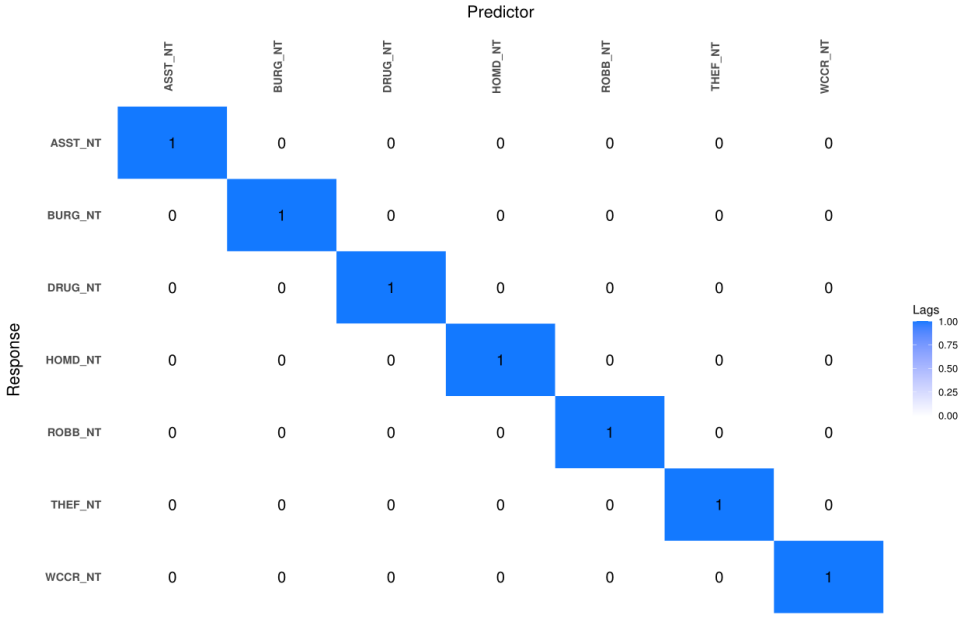
Transforming the data to make it stationary was a necessary first step, considering non-stationary data cannot be used in forecasting and modeling. We applied a difference and log difference transformation to the data (each of the crime types) to stabilize the mean and variance of each time series.

After the data was transformed, we utilized the bigtime, vars, and other packages in the RStudio environment to fit a SVAR model to the data. We split the data using a train test ratio of 80/20. The SVAR model using the “HLag” criteria produced the lowest test error on the TAAG and non-TAAG data. This result was great because Professor Shin and her previous student Ivan did research and found the method of sparse var w/ HLag penalty with the expectation that this model would perform well for our data which we were able to confirm when compared to a regular var and AR (1) model.

TAAG lag matrix heatmap



NON-TAAG lag matrix Heatmap



The heatmap for the results of the SVAR model is pictured above.

Some additional goals of the project evolved throughout the semester. These include:

- Applying a heatmap to the Dallas County area with a slider to highlight how crime counts in each TAAG area changed over time (I.e., how did the total number of assaults change year over year from June 2014 to December 2020). Unfortunately, we were unable to get this working in the end.
- Apply additional modeling techniques such as traditional VAR (non-sparse), AR models (modeling by crime type) and evaluate the accuracy of the models using the root mean square error of the predictions. This error technique is similar to the technique detailed in the document titled “NicholsonEtal2020.pdf” within the References folder in Box.
- Predict crime values for specific weekly time periods in the data and compare the forecasted values with the observed values.

FUTURE WORK

If the project were to continue for another semester, I would want to be able to furnish a web application that could be used as a resource for the Dallas Police Department.

In the future, we would like to see how well the SVAR modeling for TAAG and non-TAAG performs by summing the forecast results and comparing these results to actual observations present in the overall dataset. Currently, we have evaluated the model on the TAAG and non-TAAG datasets, respectively. This would provide a more comprehensive evaluation of the final model that we selected.

We would also like to see how our models perform by removing the taag data for taags that are added after we start training the model. This way we could get a better understanding of the data and make a valid conclusion from our data. Because as we have it right now there are maybe two or three taag areas that become taag areas part way through. This means that the data for a taag area previously is in the non-taag dataset and in the taag dataset after it was added as a taag. This leads to a curruption of our data and models.

When looking at seasonality patterns and plotting general crime count over different date metrics, we made the claim that there Is a higher crime count during the summer months. We could have statistically proven this by calculating the average over the summer months and then comparing this to the average mean of the winter months to bolster our statement. This is known as taking a rolling average.

Also, when we created stacked bar plots using plotly express package in python, to give viewers a better understanding of the data, we could add the percentage of the whole for each category. This will give a better understanding of what proportion of the total each subcategory represents.

RESOURCES & SOURCES

TAAG.pptx (provided by Dr. Shin). This PowerPoint provided an overview of the project, the project goals, and the references to become familiar with the approach to completing the project. This PowerPoint supplied the historical context information that we included in our final presentation.

Software used to complete the project includes Google Colaboratory, a Python programming notebook embedded in a web browser. Commonly called Google Colab, it is a product from the Google Research team that allows anyone to write and execute python code in their web browser and provides free access to Google's computing resources. Google Chrome is the preferred browser to run code in the Colab. We also used RStudio which utilizes the R programming language. RStudio requires the download of the R language and RStudio IDE.

The project was completed on the laptop of everyone's choice using a shared Box account provided by UTD to store all project-related files in one location in the cloud. The code files, references we found useful, datasets, etc. were stored in the Box account.

All references are provided in the Appendix section. They are cited properly.

APPENDIX & REFERENCES

Research Paper References

John L. Worrall & Andrew P. Wheeler (2019) Evaluating Community Prosecution Code Enforcement in Dallas, Texas, Justice Quarterly, 36:5, 870-899, DOI: 10.1080/07418825.2018.1438497

MacDonald, J., Fagan, J., & Geller, A. (2016). The effects of local police surges on crime and arrests in New York City. *SSRN Electronic Journal*, 1–13. https://doi.org/10.2139/ssrn.2614058

Nicholson, W. B., Wilms, I., Bien, J., & Matteson, D. (2020). High Dimensional Forecasting via Interpretable Vector Autoregression. *Journal of Machine Learning Research 21*, 1–52.

Richard A. Davis, Pengfei Zang & Tian Zheng (2016) Sparse Vector Autoregressive Modeling, Journal of Computational and Graphical Statistics, 25:4, 1077-1096, DOI: 10.1080/10618600.2015.1092978

Wilms, I. (2021, August 9). Package 'bigtime'. Maastricht; CRAN.

Web References

AnonymousAnonymous 3922 bronze badges, javlacallejavlacalle 11.2k2727 silver badges5353 bronze badges, & Richard HardyRichard Hardy 51.8k1010 gold badges9191 silver badges212212 bronze badges. (2015, March 1). *How does auto.arima deal with The a leap year in R?* Cross Validated. Retrieved October 15, 2021, from https://stats.stackexchange.com/questions/133504/how-does-auto-arima-deal-with-the-a-leap-year-in-r.

Fellow, S. M. P., Lead, C. P. M. L., Practitioner, A. A., Analyst, O. C. B. D., David Morton de Lachapelle Chief Scientific Officer, Manager, O. B. P., & Specialist, C. C. S. (2021, October 21). *Deep Learning for Time Series forecasting*. Machine Learning Mastery. Retrieved December 4, 2021, from https://machinelearningmastery.com/deep-learning-for-time-series-forecasting/.

ferrelwillferrelwill 61722 gold badges55 silver badges1515 bronze badges, & PedrolarbenPedrolarben 1. (1968, August 1). *Normalize time-series data before or after split of training and testing data?* Stack Overflow. Retrieved December 4, 2021, from https://stackoverflow.com/questions/62733351/normalize-time-series-data-before-or-after-split-of-training-and-testing-data?rq=1.

Hastie, T., Friedman, J., & Tisbshirani, R. (2017). *The elements of Statistical Learning: Data Mining, Inference, and prediction* (2nd ed.). Springer.

Iordanova, T. (2021, November 29). *An introduction to non-stationary processes*. Investopedia. Retrieved December 1, 2021, from https://www.investopedia.com/articles/trading/07/stationary.asp.

lionelderkrikor. (2020, April 1). *How to back-transform differentiated time series data?* RapidMiner Community. Retrieved December 4, 2021, from https://community.rapidminer.com/discussion/57266/how-to-back-transform-differentiated-time-series-data.

Mohr, F. X. (2020, August 13). *An introduction to structural vector autoregression (SVAR)*. econometrics. Retrieved December 2, 2021, from https://www.r-econometrics.com/timeseries/svarintro/.

Olah, C. (2015, August 27). Understanding LSTM Networks [web log]. Retrieved November 30, 2021, from https://colah.github.io/posts/2015-08-Understanding-LSTMs/.

Pfaff, B. (2021, September 17). *Var modelling [R package vars version 1.5-6]*. The Comprehensive R Archive Network. Retrieved November 15, 2021, from https://cran.r-project.org/web/packages/vars/.

Rasheed, R. (2020, July 11). Why Does Stationarity Matter in Time Series Analysis? [web log]. Retrieved December 1, 2021, from https://towardsdatascience.com/why-does-stationarity-matter-in-time-series-analysis-e2fb7be74454.

Svmflower, svmflowersvmflower 11311 gold badge11 silver badge66 bronze badges, & Glen_bGlen_b 255k3030 gold badges545545 silver badges926926 bronze badges. (1963, January 1). *Time Series Forecast: Convert differenced forecast back to before difference level*. Cross Validated. Retrieved December 4, 2021, from https://stats.stackexchange.com/questions/126525/time-series-forecast-convert-differenced-forecast-back-to-before-difference-lev.

Wbnicholson. (2019, June 22). *smallmedium.r*. GitHub. Retrieved November 30, 2021, from https://github.com/wbnicholson/HLAG/blob/master/code/Application/FRED/smallmedium.R.

Wbnicholson. (2019, June 22). *smallmedium.r*. GitHub. Retrieved November 30, 2021, from https://github.com/wbnicholson/HLAG/blob/master/code/Application/SW/smallmedium.R.

Websites Used

https://colah.github.io/posts/2015-08-Understanding-LSTMs ✔

"https://www.investopedia.com/articles/trading/07/stationary.asp" ✔

https://towardsdatascience.com/why-does-stationarity-matter-in-time-series-analysis-e2fb7be74454 ✔

https://www.r-econometrics.com/timeseries/svarintro/ ✔

https://machinelearningmastery.com/deep-learning-for-time-series-forecasting/ ✔

https://stackoverflow.com/questions/62733351/normalize-time-series-data-before-or-after-split-of-training-and-testing-data?rq=1 ✔

https://stats.stackexchange.com/questions/126525/time-series-forecast-convert-differenced-forecast-back-to-before-difference-lev ✔

https://community.rapidminer.com/discussion/57266/how-to-back-transform-differentiated-time-series-data ✔

https://hastie.su.domains/ElemStatLearn/printings/ESLII_print12_toc.pdf ✔

https://github.com/wbnicholson/HLAG/blob/master/code/Application/SW/smallmedium.R ✔

https://github.com/wbnicholson/HLAG/blob/master/code/Application/FRED/smallmedium.R ✔

https://cran.r-project.org/web/packages/vars/ ✔

https://stats.stackexchange.com/questions/133504/how-does-auto-arima-deal-with-the-a-leap-year-in-r ✔

CONTACT INFORMATION

Matt Brown -

meb180001@utdallas.edu

John Kenney -

jfk150030@utdallas.edu

Kaushik Pasikanti -

kxp170004@utdallas.edu

Abdel Homi -

abdel.homi@utdallas.edu

PRINT/SIGN/DATE (pending Dr. Shin's signature)

Company mentor, faculty advisor, and each team member should read and agree by signing this document and submit an electronic version (PDF/DOC) through eLearning. *Make sure to print the full name of each person signing this document*

Matthew Brown	*Abdel Homi*
Kaushik Pasikanti	*John Kenney*
	Date: 12/03/21