Jagannathan Govindan

IST 718

Lab #2

**Introduction**

For this lab exercise we are going to look at real estate as an investment opportunity and come up with best predictor from the given dataset. Linear regression and time series analysis will be used essentially for the same. The dataset provided has several zip codes in Arkansas metro areas which will be analyzed. Using the model, there will be a comparison for the national data. The final step is to see whether zip codes can be a best predictor variable.

**Obtain and Scrub:**

Multiple datasets were used and criminal stats were appropriated from the FBI crime data.

Files.zillowstatic.com/research/public/Zip/Zip\_Zhvi\_SingleFamilyResidence.csv

<https://download.bls.gov/pub/time.series/la/>

The main dataset Zip\_Zhvi\_SingleFamilyResidence.csv was collected from Zillow, then used to estimate housing worth per city and state. This dataset covers a time series between 1996-01 through 2019-09. However, based on the other data available, some of them were trimmed for better clarity. FBI crime states are aggregated based on state county:

* maryland.xls
* new-hampshire.xls
* virginia.xls
* district-of-columbia.xls

having the crime data was helpful in getting the Zillow data frame filtered based on specific crimes in those metro areas.

**Learning:** For this lab exercise, the % crime rate allowed for a population was set to be 3% as per the inputs from our team. So any county that exceeds this value was omitted from the data aggregation. Next labor force data was collected from the Department of Labor:

* laucnty17.xlsx

When joined with the latter combined data, the composition allowed further filtering based on unemployment rate. For this study, unemployment was not allowed to exceed 3.5%. Finally, the adjusted data was filtered on select locations to reduce the modeling:

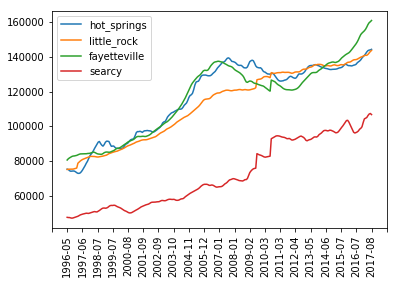
* Maryland
* Virginia
* Washington D.C.
* Massachusetts
* New Hampshire

**Learning:**

Used the zipcodes package to have zipcode column and merged them appropriately to be part of the dataframe.

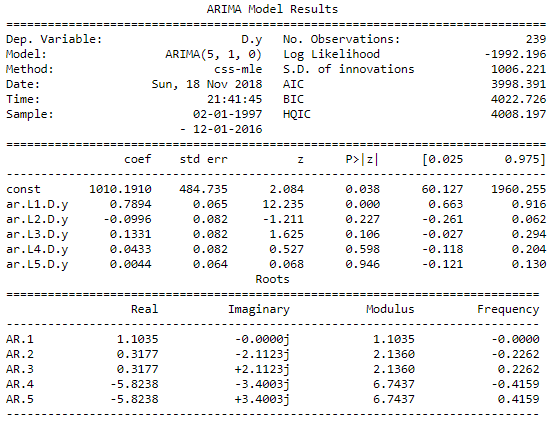
**Results**

Time series models were generated for metro areas in Arkansas:



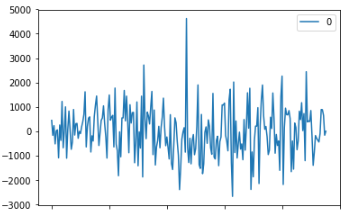
Time series plots for metro areas in Arkansas from Zillow dataset.

It appears that Fayetteville, Arkansas has the greatest increase of property value, where Little Rock and Hot Springs are closely entwined. Next, an ARIMA model was generated using a train dataset to determine whether a time series model could generalize housing data. This was done from 01/1997 to 01/2017:

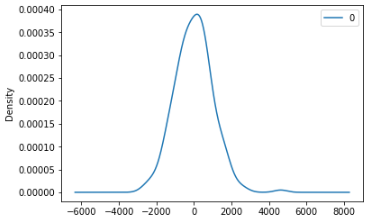


Descriptive statistics for an overall ARIMA model between 01/1997 and 01/2017 (train set).

An overall residual (Figure 3), and kernel density estimation (kde) plot (Figure 4) were generated using the same train dataset:

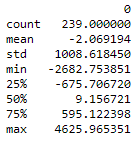


Residual plot for an overall ARIMA model.



KDE plot for an overall ARIMA model.

Descriptive statistics for the overall ARIMA model:



Descriptive statistics for the overall ARIMA model.

The predicted values with error were computed for the remaining months in 2017, which were not in the train dataset as well as for the 2018 months:

=======================================

date: 2017-2

---------------------------------------

predicted=388813.576701, expected=391050.000000

prediction difference: 0.005719

=======================================

predicted=388813.576701

=======================================

date: 2017-3

---------------------------------------

predicted=390443.689396, expected=392900.000000

prediction difference: 0.006252

=======================================

predicted=390443.689396

=======================================

date: 2017-4

---------------------------------------

predicted=392003.290988, expected=395800.000000

prediction difference: 0.009592

=======================================

predicted=392003.290988

=======================================

date: 2017-5

---------------------------------------

predicted=393503.316512, expected=398850.000000

prediction difference: 0.013405

=======================================

predicted=393503.316512

=======================================

date: 2017-6

---------------------------------------

predicted=394958.772728, expected=400850.000000

prediction difference: 0.014697

=======================================

predicted=394958.772728

=======================================

date: 2017-7

---------------------------------------

predicted=396374.708561, expected=402250.000000

prediction difference: 0.014606

=======================================

predicted=396374.708561

=======================================

date: 2017-8

---------------------------------------

predicted=397752.815045, expected=403400.000000

prediction difference: 0.013999

=======================================

Test MSE: 22164768.684165

ARIMA prediction and error rate, from the remaining 2017 months not in train.

---------------------------------------

predicted=409223.452509

=======================================

date: 2017-10

---------------------------------------

predicted=413533.867822

=======================================

date: 2017-11

---------------------------------------

predicted=417515.922014

=======================================

date: 2017-12

---------------------------------------

predicted=421472.969033

=======================================

date: 2018-1

---------------------------------------

predicted=425233.650838

=======================================

date: 2018-2

---------------------------------------

predicted=428722.288873

=======================================

date: 2018-3

---------------------------------------

predicted=431986.176158

=======================================

date: 2018-4

---------------------------------------

predicted=435068.894530

=======================================

date: 2018-5

---------------------------------------

predicted=437986.877565

=======================================

date: 2018-6

---------------------------------------

predicted=440749.383756

=======================================

date: 2018-7

---------------------------------------

predicted=443369.575329

=======================================

date: 2018-8

---------------------------------------

predicted=445861.222357

=======================================

date: 2018-9

---------------------------------------

predicted=448234.798712

=======================================

date: 2018-10

---------------------------------------

predicted=450501.831974

=======================================

date: 2018-11

---------------------------------------

predicted=452672.004600

=======================================

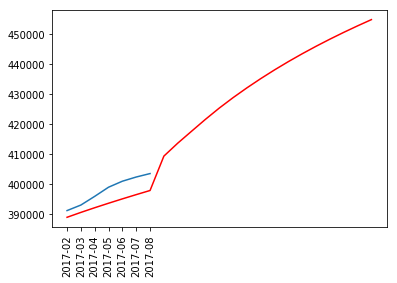
date: 2018-12

---------------------------------------

predicted=454753.750478

ARIMA predictions for 2018 months..

To visualize the difference between the predicted and actual values:



**Figure 8**. Comparison of the predicted (red) vs. actual mean (blue).

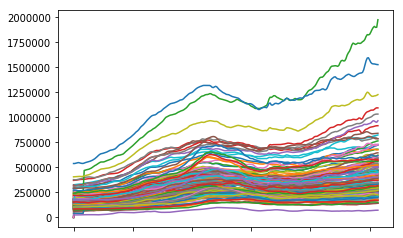
Upon reviewing the above train case, it is evident that none of the difference between the predicted with the actual test data exceed 1.5%. Therefore, the implementation of ARIMA(5,1,0) for the entire population, was extended for each individual zip code. However, since the internal statsmodels.tsa.arima\_model does not allow d > 2[[1]](#footnote-1), the differencing was indirectly applied, by adapting the actual dataset only when the base implementation would not succeed.

A computed rolling prediction successfully generated n predictions after the train. Conveniently, earlier edge cases did not arise. Therefore, each computation succeeded with a standard non-iterative differencing strategy.

predictions: [441994.60672424873, 447533.39722298674, 452893.74103927833, 457767.27660728176, 462337.7806300529, 466176.1912943393, 468413.75085113844, 470550.8847704935, 473187.8653109091, 475525.0143412446, 478662.4989835472, 482899.99798691145, 487237.503724659, 491175.01066866284, 494912.5144486477, 499150.0146820069, 504687.51469225716, 511125.0146084265, 517462.5145216945, 523000.0144826223, 527837.5144814998, 532075.0144836435, 535212.5144869951]

Rolling prediction for a given zipcode in the aggregated list.

An aggregated plot of the full original dataset duration indicates similar trends:

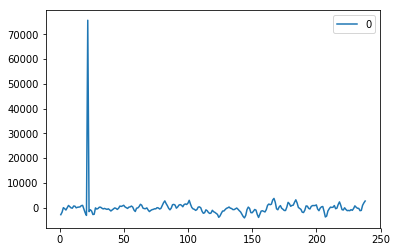


Zipcode trend for select zipcodes.

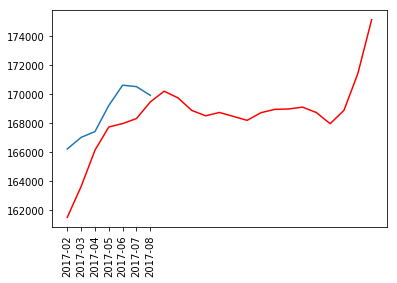
As this plot shows the difference between zipcodes is very minimal, so trying the model over multiple iterations might not be beneficial. However it can be better smoothened. After models were generated from the train, 23 iterative months were predicted. The first seven months of this iterative process coincided with the test data and was used to validate the strength of the model. The successive 16 months were used for future forecasting.

The most accurate model was found to be Grottoes, VA (24441). The rolling prediction for successive months starting 2017-02:

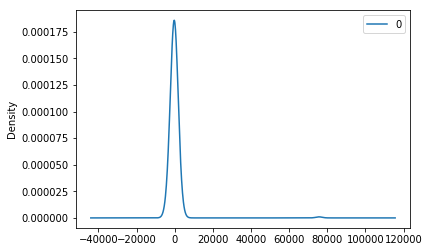
1. $161,484.18/0.01
2. $163,610.26/0.006
3. $166,112.94/0.005
4. $167,715.31/0.007
5. $167,951.84/0.01
6. $168,296.66/0.01
7. $169,441.690.008
8. $170,185.82
9. $169,725.05
10. $168,858.37
11. $168,488.13
12. $168,717.51
13. $168,446.43
14. $168,174.83
15. $168,702.87
16. $168,930.72
17. $168,958.52
18. $169,086.27
19. $168,713.98
20. $167,941.67
21. $168,869.34
22. $171,397.01
23. $175,124.68



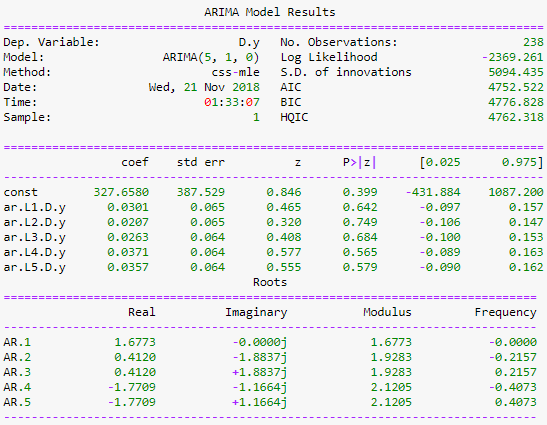
Residuals plot for most accurate model (24441).



Most accurate model (red), against actual (blue) (01002).



KDE plot for most accurate model (24441).

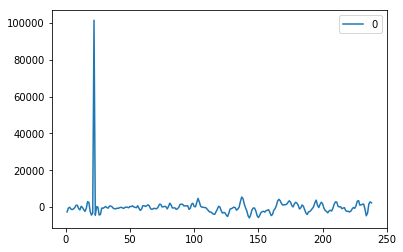


Model summary for most accurate model (24441).

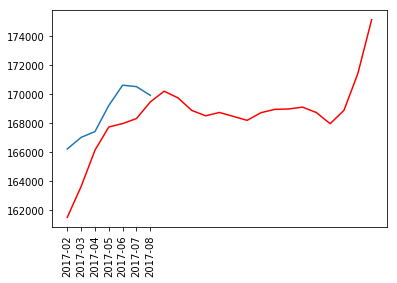
The second most accurate model was found to be Dayton, VA (22821). The rolling prediction for successive months starting 2017-02:

1. $198,827.60/0.01
2. $198,677.87/0.0006
3. $199,964.44/0.002
4. $201476.82/0.007
5. $201,596.99/0.006
6. $202,217.43/0.01
7. $203,639.40/0.02
8. $205,962.23
9. $206,785.41
10. $204,108.72
11. $201,832.09
12. $204,355.50
13. $208978.92
14. $212,402.35
15. $215,625.79
16. $219,249.22
17. $221,972.66
18. $222,996.10
19. $224,019.54
20. $224,742.97
21. $224,466.41
22. $224,489.85
23. $225,413.29

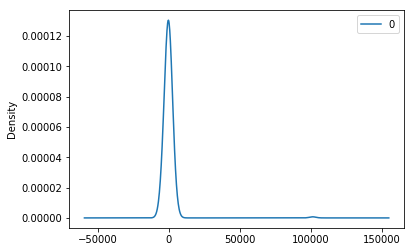
**Note:** the first seven months (test) above, provide a prediction significance against the corresponding test value.



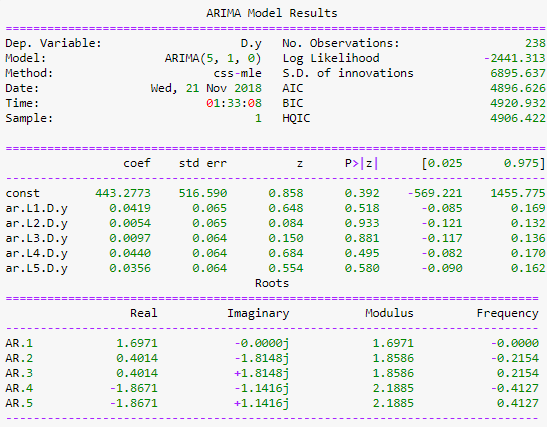
Residuals plot for most accurate model (22821).



Most accurate model (red), against actual (blue) (22821).



KDE plot for most accurate model (22821).

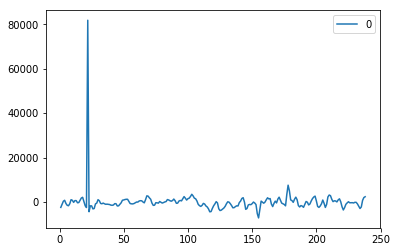


Model summary for most accurate model (22821).

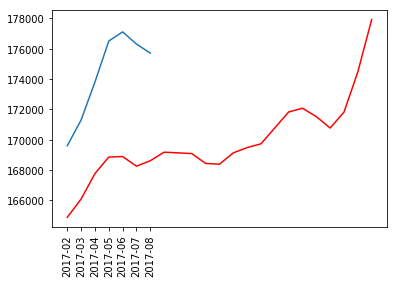
The third most accurate model was found to be Elkton, VA (22827). The rolling prediction for successive months starting 2017-02:

1. $175,100.62/0.03
2. $174,356.88/0.01
3. $174,299.22/0.002
4. $175,067.27/0.008
5. $175,429.34/0.009
6. $175,691.15/0.003
7. $176,753.34/0.005
8. $177,813.93
9. $178,073.80
10. $177,533.66
11. $176,793.45
12. $177,853.19
13. $180,512.94
14. $183,972.70
15. $187,432.47
16. $190,192.23
17. $192,251.99
18. $195,111.76
19. $198,171.53
20. $199,131.29
21. $198,691.06
22. $198,450.82
23. $198,410.59

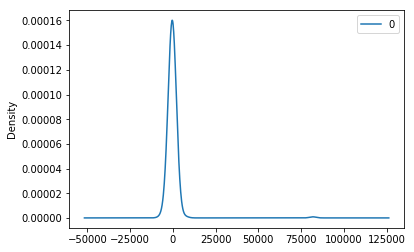
**Note:** the first seven months (test) above, provide a prediction significance against the corresponding test value.



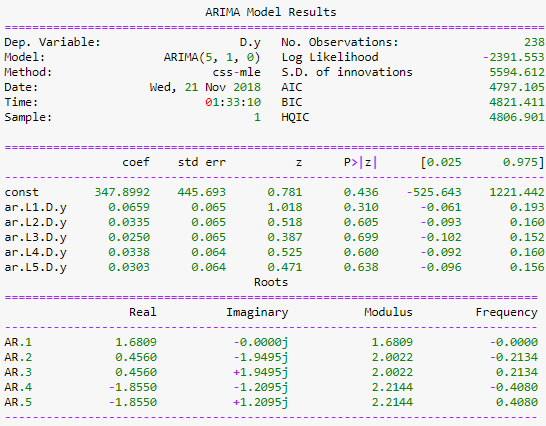
Residuals plot for most accurate model (22827).



Most accurate model (red), against actual (blue) (22827).



KDE plot for most accurate model (22827).

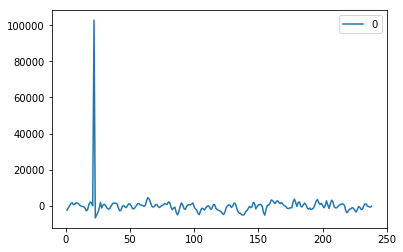


Model summary for most accurate model (22827).

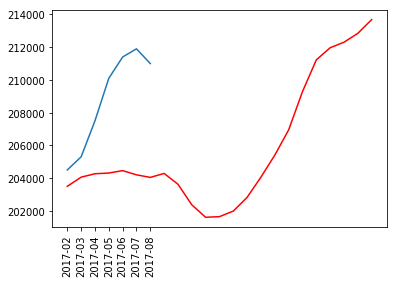
The fourth most accurate model was found to be Ringe, NH (03461). The rolling prediction for successive months starting 2017-02:

1. $207,893.76/0.01
2. $207,011.27/0.008
3. $206,969.19/0.002
4. $207,273.43/0.01
5. $208,090.60/0.01
6. $209,303.86/0.01
7. $210,618.83/0.001
8. $212,134.11
9. $214,449.07
10. $216,363.91
11. $217,078.79
12. $217,393.64
13. $217,908.49
14. $218,723.33
15. $219,938.17
16. $221,153.02
17. $222,367.86
18. $224,982.70
19. $227,997.55
20. $229,712.39
21. $230,627.24
22. $230,142.08
23. $228,856.92

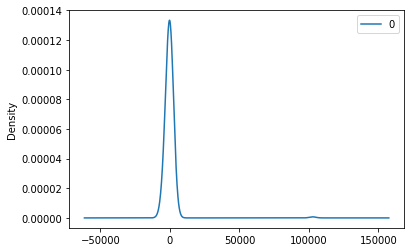
**Note:** the first seven months (test) above, provide a prediction significance against the corresponding test value.

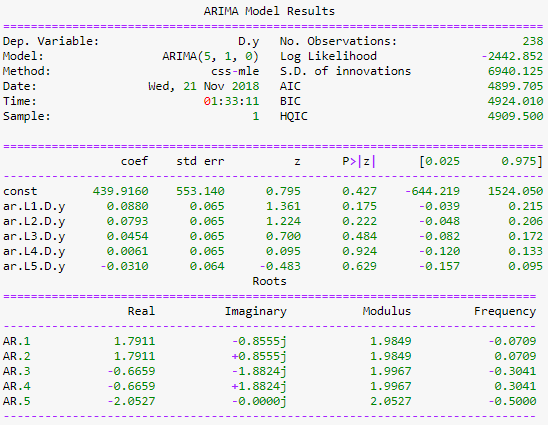


Residuals plot for most accurate model (03461).



Most accurate model (red), against actual (blue) (03461).

 KDE plot for most accurate model (03461).



Model summary for most accurate model (03461).

When reviewing the top four performing models, it is evident that the first seven months for each model, never exceeded 3% difference between from the corresponding test value. Therefore, it is safe to assume the corresponding models are valid, and the successive 16 months are interesting.

**Conclusions**

As part this exercise first, we look at the data from Arkansas Metro Area and how the prediction models come up. Then tried, an overall ARIMA rolling model which proved to be a highly accurate model. Specifically, at each interval-step, the predicted value contained less than a 1% error. Therefore, this approach was adopted while merging dataset by zipcode, then predicting successive months of prediction.

When reviewing the above results, it was found that the top four zipcodes aggregated by crime, unemployment, and by location (MD, VA, D.C., MA, NH) were three in Virginia, and one New Hampshire. From the descriptive statistics, the model generated an accurate results.

Python code:



1. [↑](#footnote-ref-1)