Predicting ZORI using COVID-19 and Census Data

RUN-PLT:
Aleksey Klimchenko, Alex Pinkerton, Nixon Lim

Introduction

Introduction

The COVID-19 pandemic drastically shifted consumer preferences in housing. This created a scenario that made housing prices difficult to model, with price variability changing constantly as the pandemic continues.

Objective

Create models that can accurately forecast and predict the Zillow Observed Rent Index (ZORI) by using historical ZORI, US Census, and Covid-19 data.

Predictions can be used to see which areas are recovering from Covid more quickly. This kind of information can be useful to:

- Government Organizations: find areas in need of resources
- Investors: find which neighborhoods to start businesses in and which to stay away from
- Business Owner: gain insight into your own neighborhood to decide when to increase or decrease inventory

Datasets

Datasets

Zillow Observed Rent Index	US Census Data	COVID-19 Data from Johns Hopkins
Monthly rent prices of the center 20% (40th-60th percentile) of the entire observed market Smoothed Seasonally adjusted	Yearly data collected and/or predicted by the US Census Bureau 2014-2018 datasets were pulled from BigQuery	Data collected by the Johns Hopkins Center for Systems Science and Engineering (JHCSSE) Daily counts of US COVID-19 cases and deaths, by US county
Jan. 2014 - June 2021 1743 rows by 93 cols	33,120 rows by 232 cols 5 years of data	Jan. 22 2020 - July 17 2021 3342 rows by 555 cols

Dataset Assumptions

Zillow Observed Rent Index	COVID-19 Data from Johns Hopkins University
• Monthly Rent Indexes (ZORI) were within the surrounding months' ZORI • Any monthly ZORI missing from the beginning of 2014 and end of 2021 followed that year's trend in ZORI change • Each ZIP code with ZORI belongs to the same county as that listed in Zillow's Home Values dataset* for that ZIP code *ZIP codes in ZORI did not have counties assigned	• COVID-19 data for each county is accurate • ZIP codes' COVID cases and deaths are proportional to the ZIP codes' population out of the total county population said ZIP code resides in

Zillow Observed Rent Index Dataset

ZORI Methodology:

- “A smoothed measure of the typical observed market rate rent across a given region… The index is dollar-denominated by computing the mean of listed rents that fall into the 40th to 60th percentile range for all homes and apartments in a given region...”
- “...smoothing is applied using a three-month exponentially weighted moving average.”
- “...smoothed indices are checked against a set of heuristics based on statistics of the time series to flag potential data quality issues...”

ZORI Cleaning

Year	Nulls	PercentNull
2014	1805	0.086667
2015	669	0.031667
2016	369	0.017500
2017	93	0.004167
2018	57	0.000833
2019	61	0.000833
2020	93	0.002500
2021	120	0.011667

Missing Values

- Interpolated from 2014-01 to 2021-06
- Any RIs missing from the beginning of 2014 or end of 2021 were calculated using that year's mean change in RI

Outliers

- Each ZIP code was checked for any RI that was >3 stddevs away from that year's mean rent price
- No outliers were found, and no ZIP codes were removed

Census Data

Cleaning: 2014-2018

- 165600 observations (33,120 ZIP codes x 5 years of data)
- 232 Features

	RegionName	year	aggregate_travel_time_to_work
4620	60602	2014	NaN
4621	60602	2015	NaN
4622	60602	2016	NaN
4623	60602	2017	NaN
4624	60602	2018	NaN

Missing Data: 2014-2018

- Removed 4 Features that had 5,000+ null values and little to no correlation with the target feature.
- Filled null values: interpolated by year within the each ZIP code
- When a ZIP code was missing all values for a feature, we used data from the closest ZIP codes to fill the null values

Multicollinearity

- Handled using Principal Component Analysis (PCA)

Census Data Cont.

Cleaning: 2019

- Pulled matching features within BigQuery Census Data from 3 tables from the Census website
- Combine features to make sure they matched with the features in the 2014-2018 dataset

Combining Census Years:

- Could only pull 47 matching features within the datasets

Multicollinearity

- Used PCA to deal with this

Feature Selection

- Reduced to 36 Features using a Forward Stepwise function

Ended with 13,920 observations (1740 zip codes*8 Years) and 36 Features

COVID-19 Data

The JHCSSE provides a daily count of global COVID-19 cases and deaths from 01/22/2020 to present day.
This data was very clean, but needed to filtered:

- Removed all observations in non-US counties
- Capped data to be from 01/22/2020 to 06/30/2021
- Cumulative numbers of cases and deaths were the only metrics kept alongside attributes

The count of cases/deaths on the last day of each month was used as that entire month's case/death count

COVID-19 Data cont.

Feature Engineering

- ‘New Cases’ and ‘New Deaths’ features:
  - Subtracted current month’s cases & deaths from previous month’s
- COVID cases/deaths for each ZIP code:
  - Assumed that the COVID cases and deaths were proportional to each ZIP Code’s population within the county
  - Integer divide to find the cases/deaths by ZIP code
  - Used 2019 Census data to find ZIP code and county -level populations
- ‘New Case Rate’ feature:
  - New cases per population of ZIP Code
- New Death Rate, Cumulative Case Rate, Cumulative Death Rate added, but not used in our final model

Combining the Datasets

1. **Combined COVID-19 and Census datasets: COVID+Census**

- Joined COVID-19 data with the Census data on the county & state
- COVID-19 case/deaths were split to the ZIP code level based on the (population of ZIP code) / (total county population)

2. **Combined ZORI dataset with the new COVID+Census dataset**

- 103 ZIP codes were not included in the SARIMA and SARIMAX models due to the lack of available corresponding Census/COVID data

Methods and Models

Our Approach

The following models were implemented to achieve our objective:

1. Multiple Linear Regression (**MLR**) to predict the yearly RI in 2022
2. Vector AutoRegression (**VAR**) to forecast the RI from 2020 to mid-2021
3. Seasonal Autoregressive Integrated Moving Average (**SARIMA**) to forecast RI in the first half of 2021
4. Seasonal AutoRegressive Integrated Moving Average with eXogenous regressors (**SARIMAX**), using COVID-19 data as the exogenous variables, to forecast RI in the first half of 2021

Methods and Models

Model	Features Used	Target
MLR	ZORI (yearly), Census 2014-19	ZORI (yearly) in 2021 and 2022
VAR	ZORI (monthly) from 01/2014 to 12/2019	ZORI (monthly) from 01/2020 to 06/2021
SARIMA	ZORI (monthly) from 01/2014 to 12/2020	ZORI (monthly) from 01/2021 to 06/2021
SARIMAX	ZORI (monthly) from 01/2014 to 12/2020 COVID-19 (monthly) from 01/2020 to 12/2020	ZORI (monthly) from 01/2021 to 06/2021

Multiple Linear Regression

Using the available data, created models to predict:

- The current rent price
- The rent price in 1 year while including the current rent price as a feature
- The rent price in first 6 months of 2021 with the last 6 months of 2020 as the test set, using a biannual aggregation
    - Training on 1/2014 - 6/2020
    - Training on 1/ 2019 - 6/2020

MLR Scores

The first 4 models were used to gauge how well MLR could predict the ZORI. This give us a basis to compare the pandemic models to.

The first 2 models shows the capability of the Census and Covid data to predict the ZORI on its own.

We, then, included the current ZORI and forecasted for the ZORI in one year. This gave us a much worse score, as expected. The model was trained on a time period without Covid to predict on a period with Covid.

It would obviously result in a bad model.

Model Attributes (yearly)	**Test Scores**
Current Rent Price - 2019	0.5761
Current Rent Price w/ PCA - 2019	0.7304
Rent Price in 1 year - 2020	0.6688
Rent Price in 1 year w/ PCA - 2020	0.6602

MLR Scores Cont.

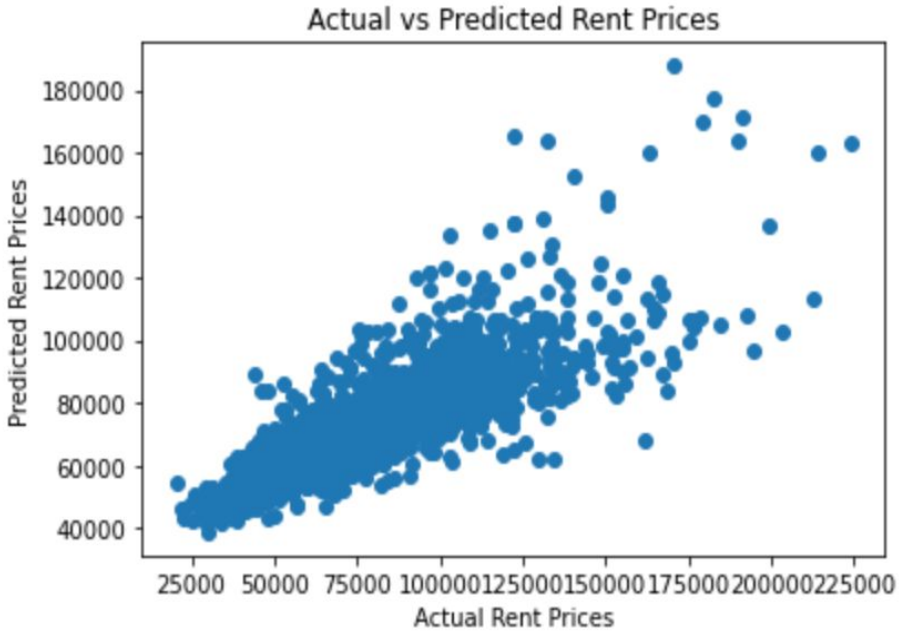
With a solid base, we had a reference to compare the next 4 models to. The first 4 model's used annually aggregated data. The data was redone to aggregate biannually. This was done to utilize the ZORI from 1/2021 - 6/2021, as the pandemic data was limited to the past 18 months.

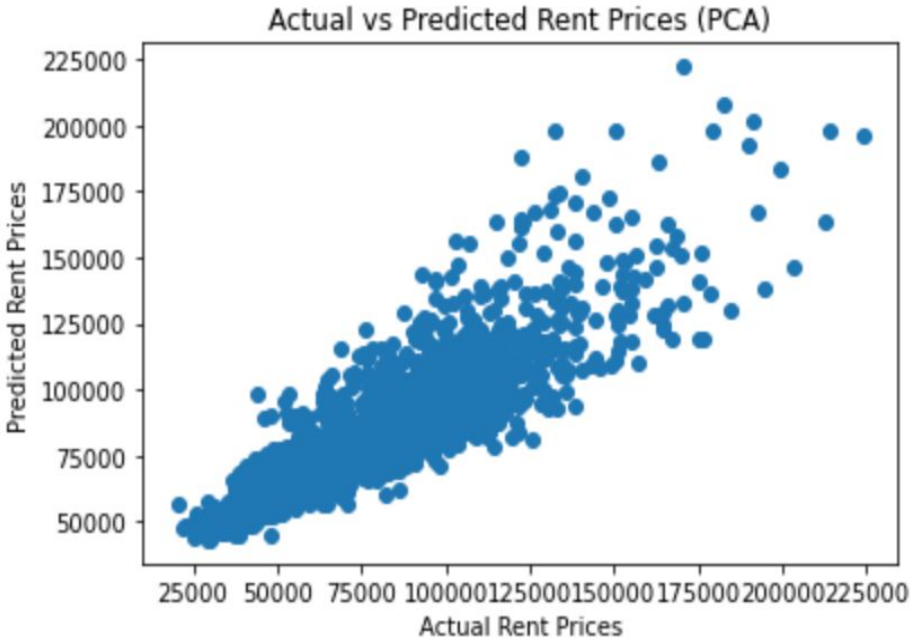
The difference between the first 2 and last 2 models are the training set. When limiting the data set to observations within the pandemic, our model did much better.

Non-pandemic data seemed to skew the data away from the actual values.

Model Attributes (biannually)	**Test Scores**
Rent Price in 6 months - 2020 (Pandemic)	0.9593 (train on 1/2014-6/2020)
Rent Price in 6 months w/ PCA - 2020 (Pandemic)	0.9593 (train on 1/2014-6/2020)
Rent Price in 6 months - 2020 (Pandemic)	0.9814 (train on 1/2019-6/2020)
Rent Price in 6 months w/ PCA - 2020 (Pandemic)	0.9966 (train on 1/2019-6/2020)

Current Rent Price - 2019





Rent Price in 1 year - 2020

Rent Price in 6 months - 2020 (Pandemic)

Trained on 1/2014 - 6/2020

Rent Price in 6 months - 2020 (Pandemic)

Trained on 1/2019 - 6/2020

VAR Model

- VAR Model was run on all 1743 ZORI ZIP codes as a ‘baseline’ time series forecast
- VAR Model proved to not be a great predictor based on RMSE scores
- However, it was a better model overall than ARIMA, ARMA, and VARMAX

VAR Results

VAR	Mean
mse	17300.830813
rmse	102.830692

VAR	zip	mse	rmse
903	60074	4.910415	2.215946
869	55117	6.752203	2.598500
1726	98270	6.972473	2.640544
121	10458	8.314598	2.883504
332	23223	9.760742	3.124219
...	...	...	...
1588	94121	415071.319614	644.260289
1382	90265	427717.908281	654.001459
120	10280	523694.218037	723.667201
117	10069	543433.510874	737.179429
1384	90272	887285.070734	941.958105

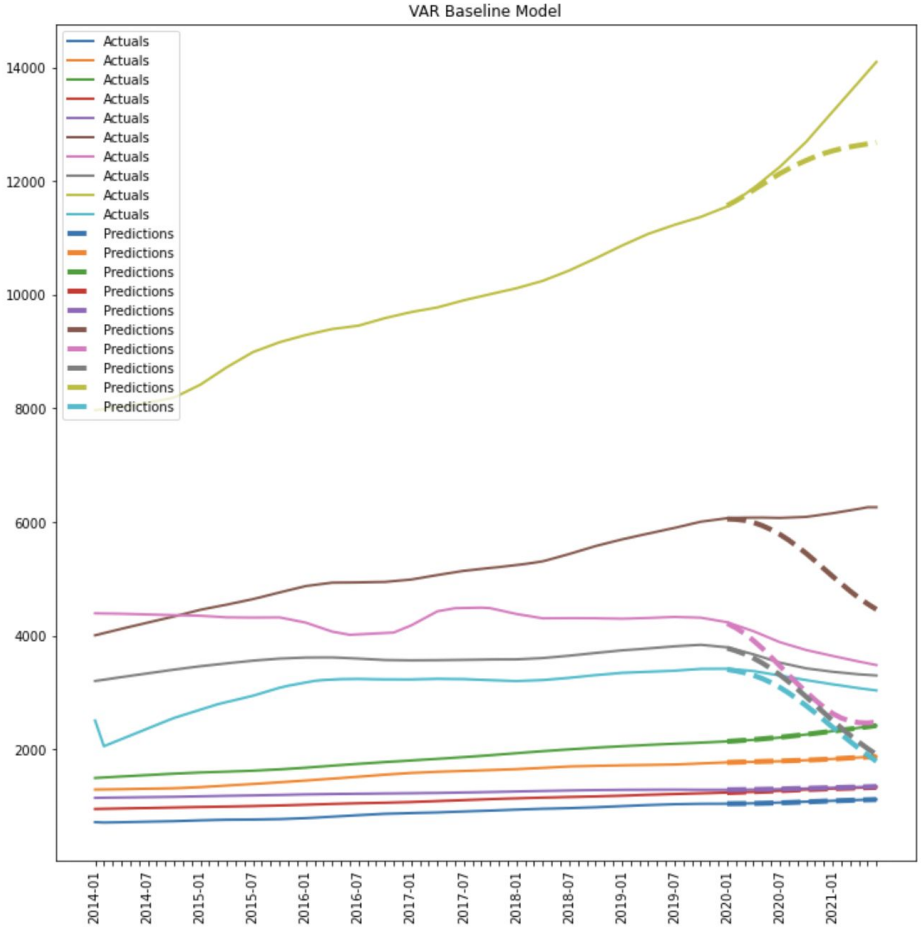
VAR Results

Looking at the best RMSE's and 5 worse RMSE's gave us a list of ZIP codes to highlight in later models-- can SARIMA and SARIMAX yield accurate predictions were VAR cannot?

VAR	zip	mse	rmse
903	60074	4.910415	2.215946
869	55117	6.752203	2.598500
1726	98270	6.972473	2.640544
121	10458	8.314598	2.883504
1098	77550	10.954256	3.309721
1588	94121	415071.319614	644.260289
1382	90265	427717.908281	654.001459
120	10280	523694.218037	723.667201
117	10069	543433.510874	737.179429
1384	90272	887285.070734	941.958105

VAR Results

VAR	zip	mse	rmse
903	60074	4.910415	2.215946
869	55117	6.752203	2.598500
1726	98270	6.972473	2.640544
121	10458	8.314598	2.883504
1098	77550	10.954256	3.309721
1588	94121	415071.319614	644.260289
1382	90265	427717.908281	654.001459
120	10280	523694.218037	723.667201
117	10069	543433.510874	737.179429
1384	90272	887285.070734	941.958105



SARIMA Parameter Tuning

Grid search was performed to determine the optimal order, seasonal order, and trend parameters for SARIMA models:

- on a mean of ZORI inputs
- parameters with lowest AIC

The optimal parameters yielded an AIC of 8.0:

- Order (p, d, q) = (0, 0, 1)
- Seasonal (P, D, Q, seasonality) = (0, 1, 0, 12)
- Trend = ‘t’
    - (linear parameter controlling the deterministic trend polynomial)

The same parameters were used when modeling with SARIMAX

SARIMA Results

(All 1640 ZIP codes)

	Mean
SARIMA	
mse	10069.026069
rmse	66.655876

	zip	mse	rmse
SARIMA			
1039	78724	0.264884	0.514669
774	53212	2.289775	1.513200
80	08807	2.408266	1.551859
1205	85712	2.737893	1.654658
784	55119	2.923646	1.709867
...	...	...	...
1472	94040	221488.341484	470.625479
1478	94107	234388.115071	484.136463
1477	94105	285397.359341	534.225944
1489	94301	378200.784523	614.980312
1279	90265	457015.263433	676.029040

SARIMA Results

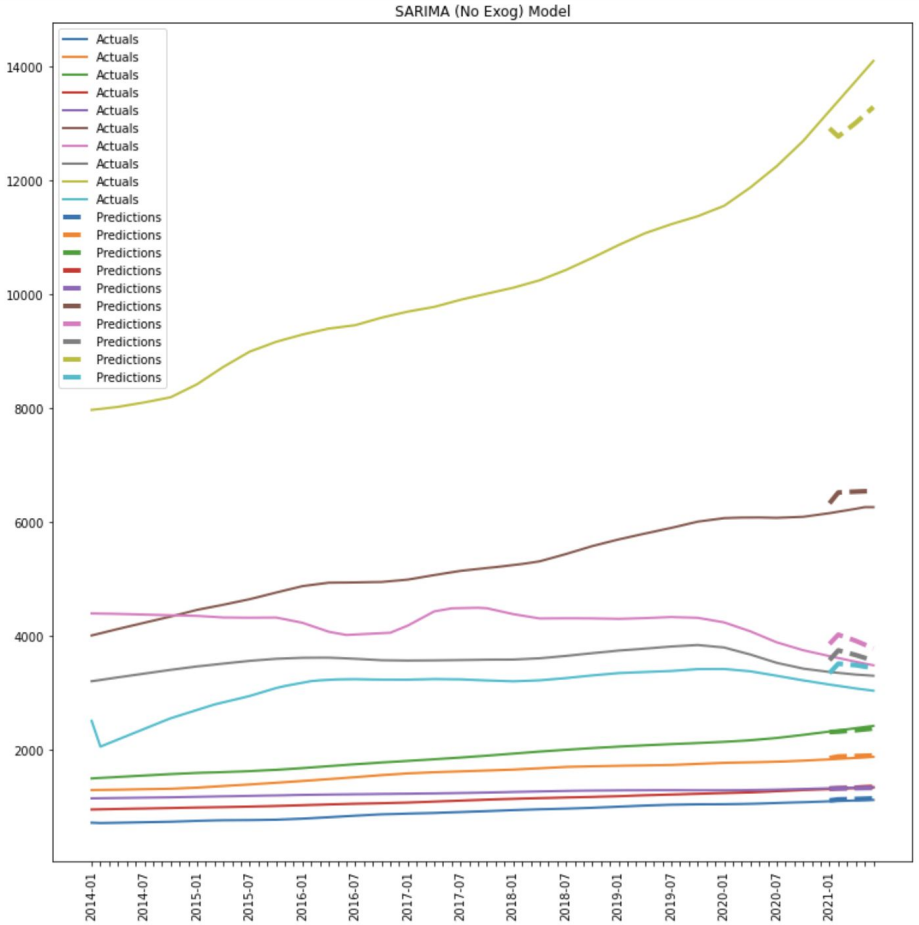
(ZIP code list from VAR)

SARIMA	zip	mse	rmse
783	55117	78.639641	8.867899
817	60074	106.101509	10.300559
995	77550	626.090108	25.021793
120	10458	1008.285191	31.753507
1623	98270	1047.994701	32.372746
1281	90272	82377.346462	287.014541
119	10280	102643.536976	320.380301
116	10069	116329.310661	341.070829
1485	94121	140261.674216	374.515252
1279	90265	457015.263433	676.029040

SARIMA Results

(ZIP code list from VAR)

	zip	mse	rmse
SARIMA			
783	55117	78.639641	8.867899
817	60074	106.101509	10.300559
995	77550	626.090108	25.021793
120	10458	1008.285191	31.753507
1623	98270	1047.994701	32.372746
1281	90272	82377.346462	287.014541
119	10280	102643.536976	320.380301
116	10069	116329.310661	341.070829
1485	94121	140261.674216	374.515252
1279	90265	457015.263433	676.029040



SARIMAX Results

SARIMAX	Mean
mse	10647.044082
rmse	69.557572

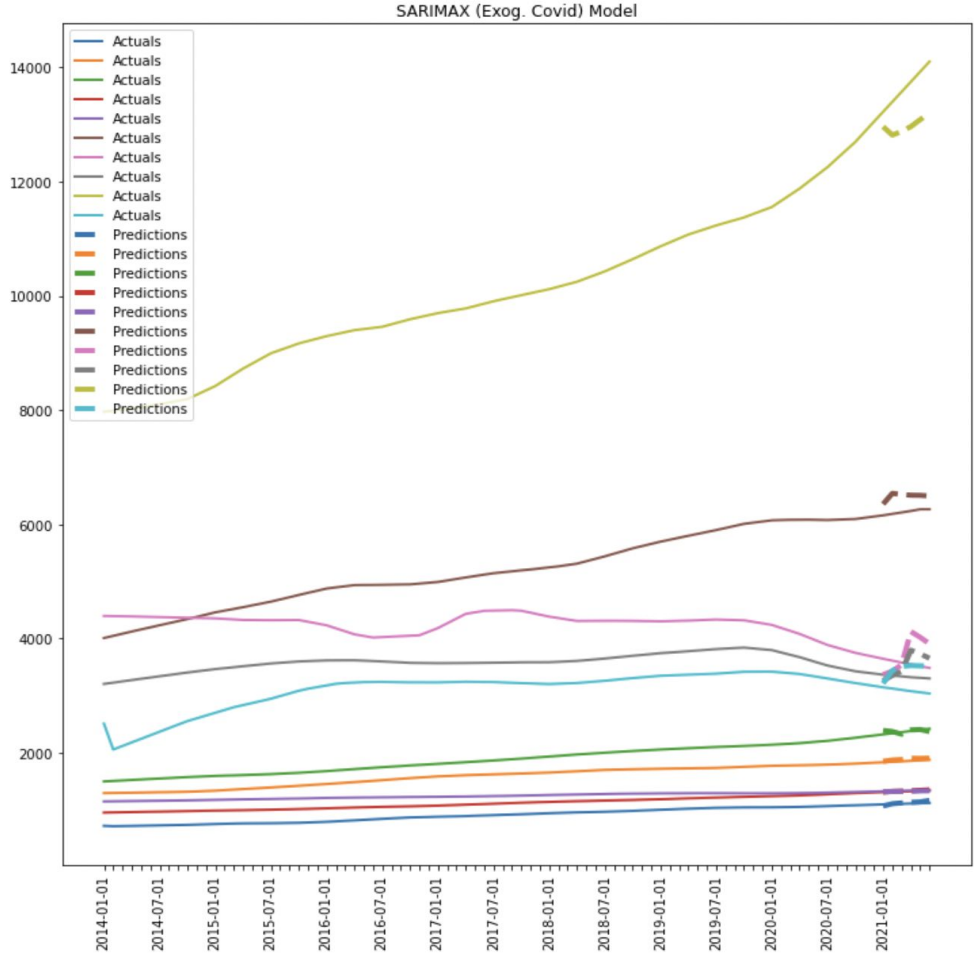
SARIMAX	zip	mse	r2	rmse
701	37138	0.522023	0.998857	0.722511
734	45324	1.018461	0.989524	1.009188
774	53212	1.103078	0.937562	1.050275
1205	85712	1.512078	0.979263	1.229666
695	37066	2.620541	0.985129	1.618808
...	...	...	...	...
1476	94103	261716.111644	-63.486363	511.581970
1478	94107	297923.358116	-66.508267	545.823560
1477	94105	374529.392880	-81.225802	611.988066
1489	94301	392672.271738	-90.803056	626.635677
1279	90265	504384.716568	-4.593647	710.200476

Sarimax Results

5 ZIP codes with the highest, lowest RMSE from VAR Model

SARIMAX	zip	mse	rmse
783	55117	78.502002	8.860136
817	60074	88.444586	9.404498
995	77550	690.126743	26.270263
120	10458	1008.708920	31.760178
1623	98270	1632.229337	40.400858
1281	90272	76005.415821	275.690797
119	10280	90862.743778	301.434477
116	10069	146202.244232	382.364021
1485	94121	153290.993562	391.523937
1279	90265	504384.716568	710.200476

Sarimax Results



Summary of Results

Model	Target	Results
MLR	ZORI (biannually) in 2021 and 2022	R^2: 0.9966 training on 2 years of past data
VAR	ZORI (monthly) from 01/2020 to 06/2021	RMSE Mean (all predictions, 18 months): $102.83
SARIMA	ZORI (monthly) from 01/2021 to 06/2021	RMSE Mean (all predictions, 6 months): $66.66
SARIMAX	ZORI (monthly) from 01/2021 to 06/2021	RMSE Mean (all predictions, 6 months): $69.56

SARIMAX using COVID-19 as Exogenous

SARIMAX models used ZIP code level COVID-19 data as exogenous factors

Results were mixed when comparing the RMSE's against those of SARIMA.

The mean of all ZIP Code's RMSE was slightly higher than that of SARIMA, but from comparing ZIP Code predictions between the two Seasonal models, an interesting pattern emerged--the ZIP Codes where SARIMAX was a better predictor were largely clustered in cities, whereas the better model predictions from SARIMA were suburban or rural.

Because the predictions are not weighted by population, and there are more suburban/rural ZIP Codes, the mean RMSE on SARIMA was slightly better

SARIMAX versus SARIMA

COVID and Census Exogenous Actual inputs must be passed when forecasting, which is a downside of the SARIMAX model.

Future predictions of Census and COVID data can be forecasted separately and then passed as the exogenous variable, but the accuracy loss of forecasting on top of forecasts would likely offset the drop in RMSE from SARIMA to SARIMAX seen in urban ZIP Codes.

Comparing SARIMA and SARIMAX RMSE

SARIMAX yielded an improved (lower) RMSE for 741 ZIP codes (45%, out of 1640):

	x_rmse	s_rmse	rmse_diff
mean	37.795993	106.646948	-68.850956
min	0.722511	8.405063	-584.927837
max	193.196816	676.029040	-0.036242

zip	x_r2	x_rmse	s_rmse	rmse_diff	CountyName	State
22203	-192.573570	148.043337	534.225944	-386.182607	Arlington County	VA
22003	0.551673	8.170123	412.801178	-404.631055	Fairfax County	VA
92021	0.839591	8.619261	438.568573	-429.949312	San Diego County	CA
98107	-219.398486	158.837982	614.980312	-456.142330	King County	WA
7047	-1097.453877	91.101204	676.029040	-584.927837	Hudson County	NJ

Comparing SARIMA and SARIMAX RMSE

SARIMAX yielded a worse (higher) RMSE for 899 ZIP codes (55%, out of 1640):

	x_rmse	s_rmse	rmse_diff
mean	95.737028	33.693267	62.043761
min	7.819605	0.514669	0.006906
max	710.200476	267.547539	700.851483

zip	x_r2	x_rmse	s_rmse	rmse_diff	CountyName	State
90265	-4.593647	710.200476	9.348994	700.851483	Los Angeles County	CA
94301	-90.803056	626.635677	14.028358	612.607319	Santa Clara County	CA
94105	-81.225802	611.988066	6.695973	605.292093	San Francisco County	CA
94107	-66.508267	545.823560	24.761957	521.061602	San Francisco County	CA
94103	-63.486363	511.581970	52.368564	459.213406	San Francisco County	CA

Comparing Models

SARIMAX for COVID and census data as exogenous factors were then input on top of the standard SARIMA inputs

Predictions and scores actually ended up further off the test actual values; downsides to this method that are worth pointing out:

COVID and Census Exogenous inputs must be passed when forecasting, so with predictions on top of predicted exogenous, forecasts can quickly become more complicated and less accurate

Further Work

SARIMAX showed an improved RMSE for many ZIP codes, leading us to believe that improving our exogenous variables we can improve our forecasts:

- Find recent/accurate population data (monthly, by ZIP would be optimal)
  - Or use 2020 census data when it becomes available
- Find accurate COVID-19 data by ZIP code
- Find other datasets that can be used for exogenous variables
  - Unemployment rates at the ZIP code or county levels
  - Dominant political party during the pandemic
  - Rates of COVID-19 recovery and/or vaccination
- Conditionally forecast 2021, 2022 ZORI based on whether COVID-19 diminishes or follows current ZIP code/county trends

Questions?