Urbanization of India

A Time Series Analysis

Nicholas V. Nikolov, Parimal Vyas

Illinois State University

Abstract

Urbanization in India has demonstrated a strong, predictable upward trend with the exception of 1981 which appeared to have been the result of fraudulent activity in which regions of India sought to increase funding by overstating population and urbanization. Several ARIMA models were compared and ultimately the strongest model predicted an urban population percentage of 33.66% with an upper tail of 36.16% for the 2021 census.

**Contents:**

Introduction

Data

Results

Data Transformation & Model Fitting

Normality

Outlier Detection

Model Diagnostic

Overfitting

Forecasting

Discussion

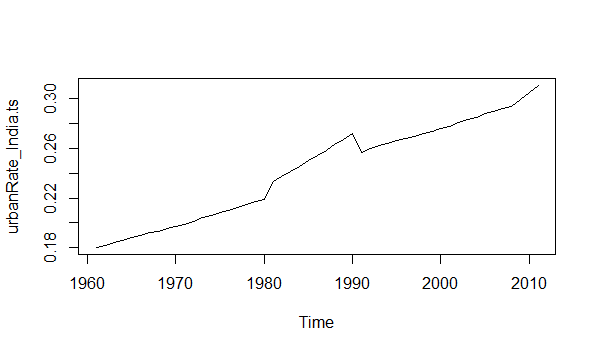
References  
Code

Urbanization of India

Urbanization is a strong indicator of the industrialization and modernization of a nation. India has demonstrated strong urban growth in recent years with the 2011 census demonstrating a slightly higher growth than predicted. The census data is collected and used to make a multitude of national decisions including funding to provinces. This project will attempt to predict future growth up to and including the 2021 census.

**Data**

Data (N=51) was collected from the UN Stats Annual Yearbook which is an annually released book of statistics on various topics pertaining to worldwide census information. The data might be presented as anything from full census to partial census to national estimates based this partial census data. In India the census is conducted every ten years on the second year of every decade (1961, 1971, 1981, etc.). Any missing years were filled using linear data interpolation so as to have a full set of data. This data interpolation was used for the years 1962, 1982, 1983, 1984, 2009, and 2010. This methodology did not affect the predictability of the model which was ultimately able to predict the 2011 urban rate within ~1% error. The time series plot below demonstrates a very strong, predictable upward trend.

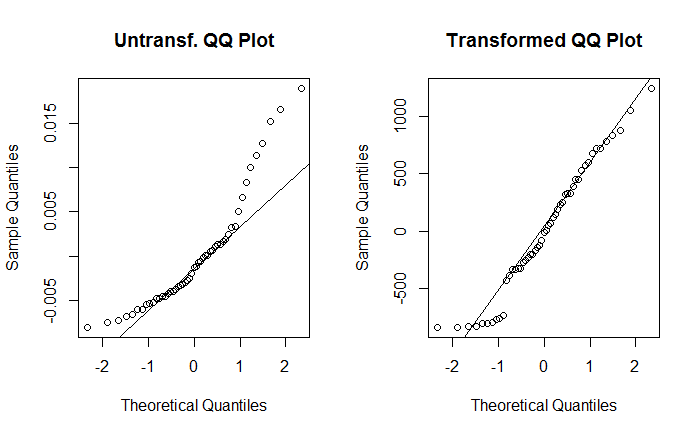


Due to a fairly small data set it was decided to split the data into a training set of 48 observations and a testing set of the remaining 3.

**Results**

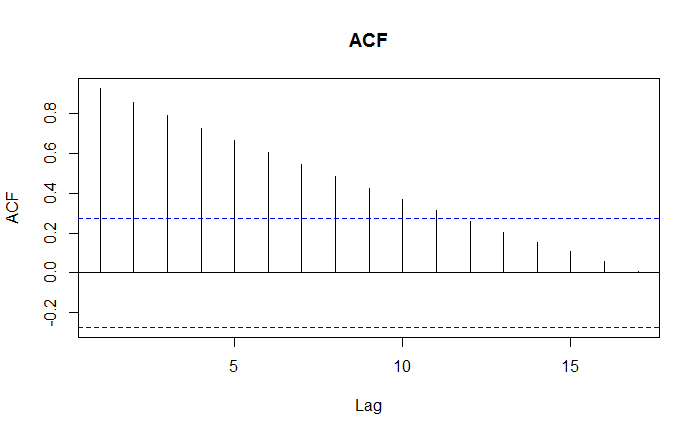
**Data Transformation**

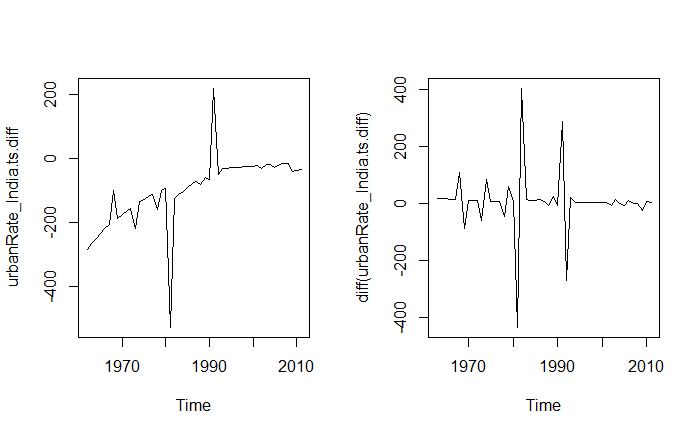
The normality assumption would allow for more accurate model selection with AIC. The residuals were analyzed and the normality assumption could not be met. The BoxCox method suggested taking one over the time series model but even that failed to meet normality expectations. Ultimately the only way to achieve normality was to take one over the Time Series Model to the fifth. The Shapiro-Wilk Normality Test demonstrated a nonsignificant p-value of 0.08336. Below is the Normal Probability Plot which shows the residuals before transformation (left) and after transformation (right).

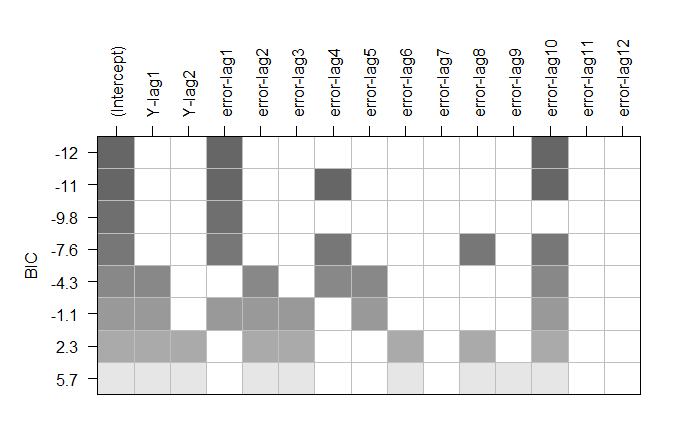


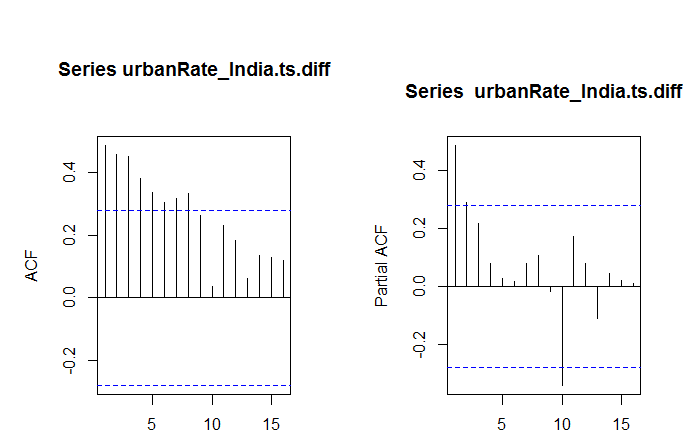
**Model Specification**

Visual analysis of the original time series model and ACF suggests nonstationarity and the Augmented Dickey-Fuller Test does, in fact, confirm that the data is nonstationary with a nonsignificant P-Value of 0.6945.

To the right can be seen the ACF for the data which strongly suggests the nonstationarity. The first difference failed to demonstrate stationarity with yet another nonsignificant P-value which required a second difference to be taken. This time the data proved to be stationary with a P-value notably less than .01. Below is a graph of the first difference (left) and second difference (right).



Now that the normality and stationarity assumptions are met we can move towards model selection. The first step taken was a quick analysis of the Best Subsets method in order to see suggested models before analyzing the ACF, PACF, and EACF plots and making manual determinations. The Subsets plot can be seen to the right. The most prevalent suggestions are MA(1) and MA(10) components.

The EACF proved rather undiscernible in that no specific model could be found. The easiest method would be to simply go with an MA(10) at this point but it is too unclear to make a valid decision. Below you will see the ACF and PACF of the data. What should be found concerning is the strong correlation spike at Lag 10. If you refer to the initial data you will find that significant outliers exist in 1981 (due to miscollection of information) and 1991 (due to return to original trend).

Ultimately it was decided that the MA(1) component would be the initial model analyzed due to simplicity. Eventually other models will be considered including, of course, the MA(10) component.

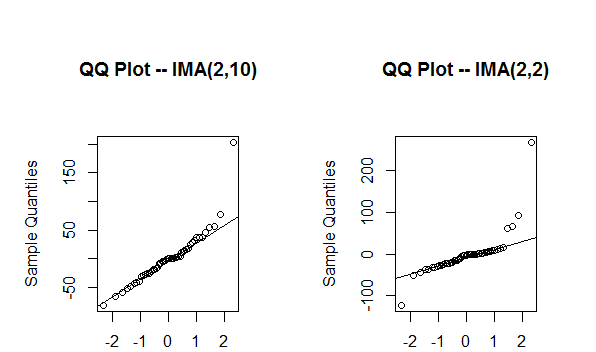
**Model Fitting and Outlier Detection**

Models are all fitted using Maximum Likelihood. Innovative Outliers were sought out and detected at observations 19, 20, 21, 30. These were incorporated into the IMA(2,1) model. No Additive Outliers were detected. Unfortunately this model presented significant dependence among the residuals as determined by the Runs Test. The Ljung-Box test suggested possible correlation among error terms. Nonetheless, let us observe our models and determine the superior one. Coincidentally, when overfitting, the IMA(0,2,2) model proved to be the most viable with the lowest AIC and BIC of all models considered. We shall compare this model to the original IMA(2,1) model to determine which model has the most accurate forecasts. Although the ma2 coefficient isn’t entirely significant (with a P-value of .0837) it will still be worth considering because it is the second strongest model.

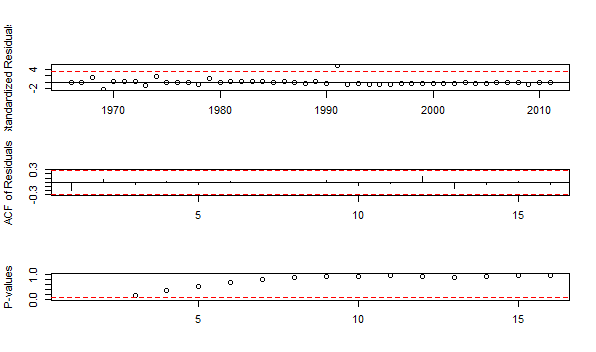
**Forecasting Set-up, Final Model Selection, and Model Evaluation**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **AIC** | **BIC** | **MSE** | **MAP** | **PMAD** |
| **IMA(2,2)** | 525.02 | 540.5426 | 9.95e-06 | 0.01033 | 0.010328 |
| **IMA(2,1)** | 547.12 | 560.71 | 1.008e-05 | 0.01040 | 0.010397 |
| **IMA(2,10)** | 536.17 | 567.1484 | 8.3239e-06 | 0.00858 | 0.008467 |

The results above are quite fascinating and it appeared necessary to consider the values of the two superior models because they appeared quite similar. Below are the Normal Probability plots of both models. Normality of residuals fails but it is relatively expected due to a very strong outlier that could not be formally smoothed.



The IMA(2,2) is the only of the two models that can pass the Runs Test above the .01 level with a P-value of 0.0117.

 Below is the overall time series diagnostic plots for the IMA(2,2) model. This demonstrates the strong outlier in 1991 (an artificially created outlier as a result of the 1981 miscount). Otherwise, the residuals are quite random and no significant correlations can be discerned.

The IMA(2,2) model was used to predict the Test Data using the One-Step Ahead Method. Below is the result of these findings. Originally the determination was made to make only 3 predictions but the years 2009 and 2010 are the results of linear interpolation connecting underestimated UN values with the higher-than-expected Urban growth recorded in 2011. The model will likely continue to under-predict so it would be wise to consider urban percentages between the estimated value and the upper tail.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Year | Actual | Predicted | Difference | %Error |
| 2008 | .294 | .294 | .000 | .00% |
| 2009 | .300 | .296 | .004 | 1.33% |
| 2010 | .305 | .302 | .003 | .984% |
| 2011 | .311 | .308 | .003 | .965% |

**Final Forecast**

The model has the following forecast for the next ten years;

Point Forecast Lo 80 Hi 80 Lo 95 Hi 95

2012 0.3137244 0.3094593 0.3179896 0.3072014 0.3202475

2013 0.3163461 0.3101482 0.3225440 0.3068672 0.3258250

2014 0.3189677 0.3112606 0.3266749 0.3071807 0.3307548

2015 0.3215894 0.3125812 0.3305976 0.3078125 0.3353663

2016 0.3242110 0.3140297 0.3343924 0.3086400 0.3397821

2017 0.3268327 0.3155662 0.3380992 0.3096020 0.3440634

2018 0.3294543 0.3171672 0.3417414 0.3106628 0.3482458

2019 0.3320760 0.3188180 0.3453339 0.3117997 0.3523523

2020 0.3346976 0.3205083 0.3488870 0.3129969 0.3563984

2021 0.3373193 0.3222307 0.3524078 0.3142433 0.3603953

This slightly under-predicts estimates found in the World Data Bank which estimates an increased growth rate. The reason for this is that other models have incorporated smoothing techniques to deal with and eliminate the outliers caused by the 1981 census. This is valid if the 2011 census was properly conducted (which showed higher than anticipated growth). We, of course, won’t know if 2011 demonstrated higher ten year growth than anticipated or if it was the result of error. For that reason the model used in this project maintained much of the trend from before the jump in 2011.

**Discussion**

It is unlikely that the 2011 census was improperly conducted but nonetheless the model was fit to resume the previous trend. Following the 2021 census it will be interesting to see if urbanization rate increases or simply returns to the same rate as before.

The IMA(2,2) model used unfortunately failed to meet residual normality assumptions. Nonetheless, the model appears to predict relatively well. Another future implementation would be to formalize full predictions for smoothed data such that the severe outliers don’t throw predictions off and actual trends can be more visible in the period around which the outliers appear.