Summer Internship 2018 IIT Delhi

Devanshu Agarwal
Advisor: Prof. Aaditeshwar Seth, IIT Delhi

***Abstract*—In any country, the districts are very different from each other in terms of socio-economic parameters. The current development status of a district can be used in making the policies which can shape the future of the district. Census plays a major role in making these policies as it is the only parameter which can give us the status of development of the district. Census data occurs typically once in a decade and hence any policies adopted as its base can be very erroneous and outdated.**
The satellite data can be used to give us the socio-economic parameters more often than census can ever give. We relate the satellite data with socio-economic parameters at the specific time frame for which both the data was available. We also made the prediction model which gives us the socio-economic parameters quite well after processing the satellite data. These predicted socio-economic parameters can be used as a proxy for making the policies.

I. Introduction

THE development status of the district can be described from various socio-economic parameters like the main source of water, the main source of light and condition of the household. These socio-economic parameters can have a very high effect on the future development of the district as the policies are based on these parameters only. The census which the government of India conducts once in a decade is the closest development status of the district that we can get. Census has to be done for every district in India and hence its highly costly in terms of time, money and human work. This makes practically impossible for the government to have census more frequently. The policies are thus very erroneous and outdated.

The satellite data namely Nightlight, MODIS and LANDSAT can be proved as a new method of finding the socio-economic parameters at both district and village level. There was a need to go to the village level also because it can give us the better insight for the area for which policy is to be made and thus can give us more effective policy. The prediction of the socio-economic parameter from satellite data can give us the socio-economic parameters for each year. Recent advancement in machine learning and machines itself gives us the facility of processing such huge sets of data.

We make nine socio-economic parameters from census data. These parameters are found out by clustering the census data. These nine socio-economic parameters are the main source of light, the main source of water, literacy, asset, bathroom facility, fuel for cooking, the condition of household, employment (agriculture or non-agriculture) and female employment. These socio-economic parameters are used rather than the census variable in all our analysis.

To make any such type of model we first see if there exist any type of relation between this satellite data and socio-economic parameter. Then we go in building the prediction model at district and village level. We also tried to explain some anomalies in nightlight values by finding events from media articles.

Advisor:Prof. Aaditeshwar Seth, IIT Delhi ÷

II. Correlation between Socio-economic Parameters

We needed to see whether these Socio-economic parameters are related to each other in any way or not i.e. are they independent of each other or not. This kind of analysis can give us a deep valuable insight into how different our predictive model will be for each socio-economic parameters. This can be done by finding the correlation value between each pair of the socio-economic parameter. Since our socio-economic parameters are discrete values, we used Cramer's V method which is a kind of normalization over a chi^2 method of correlation.

After performing Cramer's V on every pair of the socio-economic parameters we found out that, none of the pairs is strongly correlated, although there are some which are showing the moderate correlation among themselves.

From Table I we can see that Main Source of Water, Main Source of Lightning and Condition of Household are moderately correlated among themselves giving the Crammer's V value of around 0.5 and similarly Fuel for Cooking, Employment and Bathroom Facility are also moderately correlated among themselves.

TABLE I
Correlation Table

LABEL 1	LABEL 2	CHI 2	CRAMER V
MSL	CHH	411.67	0.57
FC	EMP	352.3	0.52
MSL	MSW	325.29	0.5
FC	BF	309.97	0.49
MSW	CHH	275.26	0.46
CHH	FC	236.61	0.43
MSL	FC	226.7	0.42
MSW	FC	218.42	0.41
BF	EMP	207.06	0.4
MSL	BF	182.88	0.38
MSW	BF	187.26	0.38
MSL	EMP	171.87	0.37
CHH	EMP	176.17	0.37
CHH	BF	169.03	0.36
MSW	EMP	78.11	0.25

III. Predicting Socio-economic States at District Level.

These new Features along with the MODIS existing features are normalized by the area of districts. These

TABLE II
NEW FEATURES FROM EXISTING MODIS FEATURES

Crop Ratio	Proportion of Cropland and Natural Vegetation area in a district
Average Urban Nightlight	Intensity of Nightlight in a district contributed by urban and Built-up area
Urban Ratio	Proportioon of Urban and Built-up area in a district
Natural	Area under all Natural Vegetations and Forests including water in a district
Crop Remain Ratio	Ratio of Cropland and Natural Vegetation after removing the Natural area (calculated above) from the district area
Urban Remain Ratio	Ratio of Urban and Built-up after removing the Natural area (calculated above) from the districts area
Forest	Area under all the Forests in a district
Grass Shrubs	Area covered by Shrublands, Savannas and Grasslands

normalized and raw features are together used to predict the socio-economic parameters. From Table IV we can see that MODIS outperforms nightlight in terms of accuracy and F1 score for almost all socio-economic parameters.
We use both the nightlight and MODIS features in our final model. Since the number of features now becomes very high, it can result in making our model very heavy and its performance can be decreased. To overcome this issue, we used a rich feature extraction. Rich feature extraction use coefficient attribute and feature importance attribute to give us the rank of each feature i.e. which feature is better than what feature. After we got the ranking table our next step was to choose what all features should be used in our prediction model. For this, we make one assumption that the first k features in the ranking table will give us f1 and accuracy score more compared to any other set of k features. Our challenge was now to find k. To find k, we started by taking only the first feature of the ranking table and noted down its f1 score, then we used first two features from the ranking table and noted down the f1 score for them also. This process continues until we covered all the features. This gives us a list of f1 scores. Now, the set of features which is giving the maximum f1 score should be the features which we will need in our prediction model. But, instead of taking these features which corresponds to maximum f1 score, we go on to find the sets of features whose f1 score is or above 0.01 less than the maximum f1 score. Among these sets, we chose the set which has the minimum number of features in it. One such table of selected features for bathroom facility is given in TableIII

The result of our final model is given in Table IV. In this table we have given accuracy and f1 score of nightlight, MODIS, nightlight and MODIS together.

IV. PREDICTING SOCIO-ECONOMIC STATES AT VILLAGE LEVEL FROM NIGHTLIGHT AND MODIS DATA.

After building the model at the district level, we had gone one step further by making the model at the village level. The village level model is a very important part of the development

TABLE III
SELECTED FEATURES FOR BATHROOM FACILITY

BF under-developed			
var	rank	coeff	pvals
Croplands	1	0.0004	1.33E-08
Urban and built-up	2	-0.0201	2.31E-02
logMean	3	-1.392	7.98E-06
mod_Urban and built-up	4	-0.098	2.06E-04
UrbanRatio	5	-0.098	2.06E-04
CropRatio	6	1.919	2.76E-07
mod_grass_shrubs	7	0.5876	3.40E-02
BF moderately-developed			
var	rank	coeff	pvals
mod_Croplands	1	-2.8438	6.44E-40
Cropland/Natural vegetation mosaic	2	-0.0006	4.00E-16
UrbanRemainRatio	3	-0.966	2.57E-18
mod_grass_shrubs	4	-0.4262	5.12E-18
Natural	5	-0.0001	4.88E-13
BF developed			
var	rank	coeff	pvals
mod_Urban and built-up	1	0.3589	1.96E-02
UrbanRatio	2	0.3589	1.96E-02
UrbanRemainRatio	3	1.091	7.35E-03
logMean	4	2.8403	6.22E-05
AvgUrbanNTL	5	-0.0011	5.96E-18
CropRatio	6	-1.8164	2.53E-20

TABLE IV
PREDICTION RESULTS

	NTL	MODIS	NTL and MODIS
	f1 , acc	f1 , acc	f1 , acc
Asset	0.64 , 0.7	0.63 , 0.72	0.67 , 0.73
BF	0.61 , 0.7	0.69 , 0.78	0.71 , 0.79
CHH	0.5 , 0.56	0.62 , 0.7	0.64 , 0.71
EMP AG NONAG	0.58 , 0.61	0.65 , 0.73	0.69 , 0.77
EMP FEMALE	0.52 , 0.53	0.64 , 0.73	0.65 , 0.74
FC	0.57 , 0.6	0.72 , 0.79	0.73 , 0.8
LIT	0.55 , 0.6	0.6 , 0.65	0.62 , 0.69
MSL	0.43 , 0.51	0.6 , 0.7	0.62 , 0.72
MSW	0.52 , 0.6	0.65 , 0.74	0.67 , 0.75

model of the country, as this model is at the more granularity level and hence can increase the efficiency of policies in the future.
At the village level, we have shape files for 6 states which corresponds to 1.97 lacs villages. Although we have census data for all the 29 states of India, we are bound by shapefiles to get nightlight and MODIS data for only 6 states. One hindrance before jumping into building any predicting model was to split the data into train, cross-validation and test set. It could be easily done from random sampling technique, but our data is already very skewed and we didn't want our model to get trained on villages of one region and tested on another region. To overcome this problem, place the villages into a grid of 75x75 according to the coordinates. Grid size was chosen such that on an average 10 villages should lie in each block of the grid. Then for each block of the grid, we randomly chose 50 percent of the villages for the train set, 30 percent for validation set and 20 percent for the test set. For blocks in which the number of villages is less than 10, we store those villages separately and than randomly chose for train, validation and test set as we have done for each block.

This method gives us a very uniform split. While doing the split the status of socio-economic parameters was unknown to our method but surprisingly it divides the data uniformly in terms of socio-economic parameters also. TableV shows how uniform the villages got split in these sets in terms of socio-economic parameters. Here again, we used log of the sum of

TABLE V
SPLIT OF DATA SET FOR MAIN SOURCE OF LIGHT

set_type	Proportion of Under Developed	Proportion of Moderately Developed	Proportion of Developed
train	0.3466	0.2178	0.4356
validation	0.3455	0.2183	0.4362
test	0.3506	0.2165	0.4329

nightlight and log of the mean of nightlight rather than using sum and mean of the nightlight and the MODIS features as it is and normalized with district area along with creating new features as described in district level prediction model above. On observing data under these features, we found out that most of the values are zero or the same for all districts and hence of no use. So to find out exactly which features are of use to us, we plotted the box plots for each feature and we kept only those features which are showing any variation among them for different classes. One example of feature discarded is given in Fig. 1 and of feature accepted is given in Fig. 2.

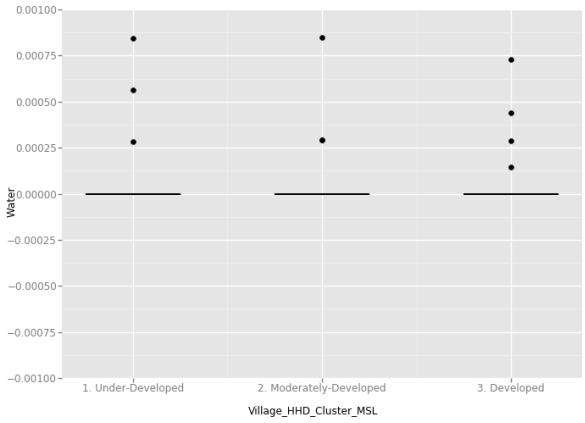

Fig. 1. Box plot of Water for Main Source of Lightning. Discared because all values of it are zero.

We then used these selected features to build our prediction model. Since our data is very huge of around 1.97 lacs villages, so there is no risk of over-fitting the data at training set as that at the district level, hence, we have used boosting algorithm (xgboost) to build the prediction model. Since our data is highly imbalanced, the minority class for socio-economic parameters were performing very poorly while the majority class was performing well. To overcome this issue, either we can do under-sampling or over-sampling. Since under-sampling would have decreased our dataset drastically, we opted for over-sampling. Synthetic Minority Over-sampling Technique or SMOTE was used to do oversampling. This

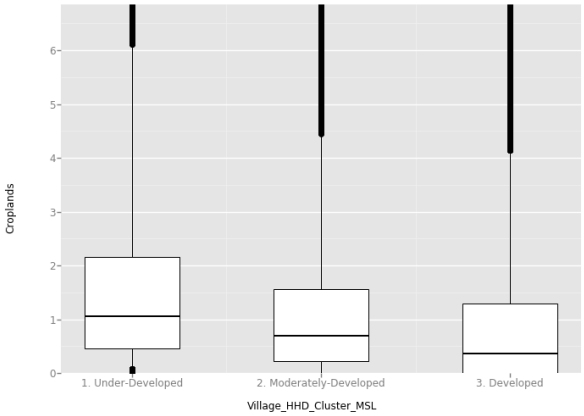

Fig. 2. Box plot of cropland. Accepted because its showing good variation.

technique generates a synthetic point rather than duplicating the point as it is. This technique is used before predicting any socio-economic parameter. The balanced class only increases the performance for minority class and it does not affect majority class much and hence the overall performance of the model does not increase significantly. Although it does not increases the performance of the model that much, still we did it because it makes our model more robust as now our model will get trained for all the classes rather than just the majority classes.

After getting balanced class we applied xgboost model with default parameters and then slowly does hyper-parameter tuning which increases our model performance fairly. Although we have addressed issues like splitting the dataset and imbalanced class quite well, the village, in general, has very less area which affects our prediction model very high as most of the MODIS and nightlight data was proved to be of no use. The results show that we can't predict at the decent level also at village level using MODIS and nightlight data and hence we have to use LANDSAT data if we have to build a decent prediction model at least.

V. PREDICTING GROWTH LABELS FROM SOCIO-ECONOMIC VARIABLES.

Alone development is not sufficient to find the state of any area. We need to find out whether that area is growing or not, so that we can have the deeper insight on that area and the policies formed can play a more effective role in the development of that area. We have tried to build the model for finding the growth status at the district level. Since we can't have census data for consecutive years, we have used the mean of nightlight as a proxy for district development and hence mean of nightlight for years 2012 to 2016 were used to find whether the district is growing or not. If the mean of nightlight was seen increasing over time then we labelled it as the growing district else non-growing district. Originally four types of clusters were observed in growth pattern but we weren't able to get a good performing model from it

and hence we limit ourselves to two clusters viz., growing or not. We called these labels, growth labels. Socio-economic variables predicted above from district-development model can be used to predict growth labels. Here also, we have removed the metro and snow-covered district as these are outliers in our case. Here also we experienced class imbalance, where a number of non-growing districts are 449 and number of growing districts are 121. The over-sampling method may have proved to be too complex to apply in this case as the independent variable in our model have discrete values and hence we opted for the simple random under-sampling method to get the balanced class. Since the independent variable, i.e., the predicted socio-economic parameters from the district-development model have discrete values in it and we have to keep our model simple because again our dataset is very small, we have used the decision tree to build our model. Decision tree along with some hyper-parameters tuning gives us the f1 score of 0.52 and accuracy score of 66 percent accuracy.

VI. Predicting Socio-economic parameters at Village level using LANDSAT data.

We have already seen earlier that nightlight and MODIS data failed to predict socio-economic parameters at village level with decent accuracy. So we have used another type of dataset in our prediction model. LANDSAT data is the satellite image of each village. Inception net was used to predict socio-economic parameters.

Two epochs were used to predict each socio-economic parameter. But we found out that increasing the number of epochs can give us better results. It takes 2 days if we want to train our model for 4 epochs and similarly increasing epochs can increase the number of days. So it is very tiresome and error-prone to change script according to epochs and labels every time we want to experiment with epochs for different labels. To solve this problem, we made one shell script along with making the given scripts more flexible so that, one can easily change the values in the shell script without affecting the python scripts directly. This gives us more freedom to play with the number of epochs and other parameters and we were able to make our results better.

As discussed earlier in predicting socio-economic parameters using nightlight and MODIS data, our dataset is highly imbalanced and hence as discussed earlier making the class balanced can be proved to be of great use. We used image augmentation as an oversampling technique. We first calculated how many more images of minority class are needed and then we augmented each image accordingly. We used random rotation as an augmentation tool.

VII. Finding media articles district wise

District and Village level models we build earlier will be used to predict socio-economic parameters. These models were build using the nightlight and MODIS data. But there can be some cases where because of some events the state of the district changes drastically. These events can by any natural calamity, election or any human organized event which includes a very large proportion of the human population involved or any other event of such huge scale. Our hypothesis was that these huge level events will influence the nightlight and the nightlight values will show an anomaly from its regular trend. If our above hypothesis is true, then we can explain any drastic change in the predicted socio-economic parameters by looking at these events. This explanation will work as support of our socio-economic parameters which our model will predict at a district level. We have not started looking into the events at village level as finding the media articles at the village level will be a challenge.

We have the mean of nightlight values for all 640 districts from April 2012 to March 2017. We also have media articles of this time span stored in our MongoDB database. The media articles stored in our database have the date attribute separately defined but not any area attribute. So we had to find out which media article is giving us information about which district if at all it is giving the district level news and not the state and India level news. Earlier we tried looking into the URL attribute and the first word of the text to find out which district the media article belongs to, but it has certain limitations like this method is not valid for some newspaper publishers of which we have media articles stored and that this method is telling us that, the media article was published in that district and it may or may not be the case that it is talking about the same district in which it got published. So we copied all the articles in which the district name we are looking for appears in the article text. We then manually read the lines in which district name is appearing in the selected articles. We found out that almost 2 percent of the cases were only wrongly classified. So we adopted this method of looking the district in the text. We then made a script which can download the media articles for any district and for any time span. One file created by the script corresponds to media articles of one day of a district. The folder and file name given by the script to these downloaded articles gives us the liberty to reuse the district and time span combination whenever required without downloading these articles again.

VIII. Finding anomalies in Nightlight.

As discussed earlier we have to first find the anomalies nightlight is showing from its trend and then we have to find out whether we can explain these anomalies by looking into the media articles or not. To reduce the affect of small changes in nightlight i.e. to make our nightlight values more smooth in terms of trend line we find out the moving average of the data for a window of 12 months and then we found anomalies on this data which is found out by subtracting the moving average correspondingly.

To find anomalies for a particular month, year and a district, we used two methods. In first method we see the growing status of nightlight for a month, year and district compared to previous month and the growing status of the same month of that district for all the other years. If the growing status of that year is opposite compared to all the other years i.e., the nightlight decreased in that month compared to the previous month for that year but for all the other years nightlight used

to get increased or vice-versa, then we can say that nightlight has shown anomaly. In second method, we first take the nightlight values for all the available years for that month and district and see how much the nightlight is showing deviation compared to other months. Deviation is found out by subtracting the average of nightlight for all the years of that month and district from the nightlight value for a month and district for the year in observation and then dividing out the standard deviation of nightlight values of all the years of that month of the district. If the deviation comes out to be greater than 1.5 than we say that the nightlight shows anomaly for that district of that particular month and year.

IX. Finding events corresponding to the anomalies found.

After finding the anomalies, we have to see whether we can explain anomalies or not by looking into the media articles. As discussed earlier we already have the script that can give us the media article for any day and hence for a month and year for any district. Since it will be very time costly to read each article manually for the whole month for a particular district and then to comment whether the anomaly can be described or not. We tried using Rapid Automatic Keyword Extraction algorithm (RAKE) to find the keywords in the given set of articles. It turns out that by looking into the keywords alone can't give us valuable insight of the set of keywords. We tried to read articles manually for few cases but it that also doesn't work for them. Moreover, we tried to search on net but we cant find anything useful for most of the anomalies labeled.
From district level we tried to go to state level and see if we can find some events at state level. To do this, we first found out how many districts in a state are showing anomalies for a particular month and a year. Then, by looking into the number of districts for each state, we selected some states which were showing anomalies for huge number of districts. For these states we again tried to find some events by applying RAKE into the media articles for all the districts of that state for that time frame and also tried to search on net for that state for that time frame but we couldn't found anything.

X. The effect of events on Nightlight.

After finding out that the anomalies marked are not showing any events from any of our models, we tried to find out how nightlight changes for a district when some huge scale events like natural disaster occur. This can give us more insight on how to find anomalies and can also give us confirmation on whether the nightlight got actually affected by events or not. The events we chose are state legislative elections in India, floods, cyclones and Kumbh Mela. By looking into these events we found out that nightlight is showing changes as expected but the deviation is very less than 1.5 chosen above. We also found out that the change in nightlight is not only affected by the event but also by the place in which that event occurs.
South Indian Flood which occured in the year 2015 affected the southern parts of India (mainly, Tamil Nadu, Andhra Pradesh and Puducherry). A plot of nighttime light values from April 2012 to March 2017 is shown in Fig 3

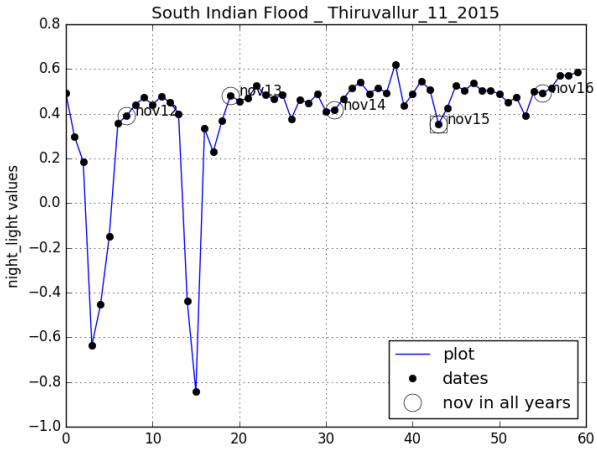

Fig. 3. South Indian Flood in Thiruvallur district

Nightlight value on November 2015 (marked with a circle inside a square) is seem to be much lower than the nightlight value in November month of 2012-2014 and 2016 years (marked with circles in a graph). Although a much downfall in nighttime light values is not seen. The possible explanation is the heavy rainfall which occurs every year in this zone of India. The average rainfall of Thiruvallur district is 1104 mm with more than 50 percent of it has been received during Northeast Monsoon Period (From October to December)
Another event we looked into was cyclone Nilam which occurred in October 2012. Mahabalipuram town in Kancheepuram district was highly affected by cyclone Nilam which occurred in the last days of October 2012. Fig 4 shows the plot of nighttime light values of Kancheepuram district from April 2012 to March 2017. Nightlight value on October 2012 (marked with a circle inside a square) seems to be much lower than the nightlight values in October month of 2013-2017 years (marked with circles in a graph) as expected.
Kumbh Mela was chosen to see the effect of it on nightlight as it is one of the biggest event organized by humans in India. It occurs periodically in Allahabad, Haridwar, Nashik and Ujjain districts. A period is of 12 years for a district. We studied nighttime light values of Nashik (Fig 6), Ujjain and Allahabad (Fig 5).

We found the significant increase in nighttime light values in Nashik and Ujjain but there is not much significant growth in Allahabad. The possible reason may be that Nashik and Ujjain are not much developed districts while Allahabad is very developed district and a lot of events occurred in it compared to Nashik and Ujjain.

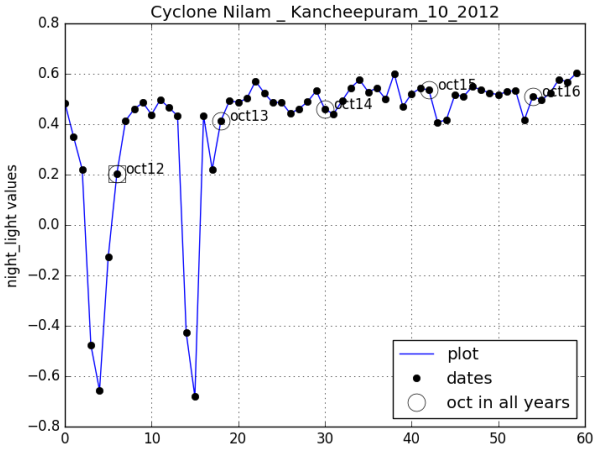

Fig. 4. South Indian Flood in Thiruvallur district

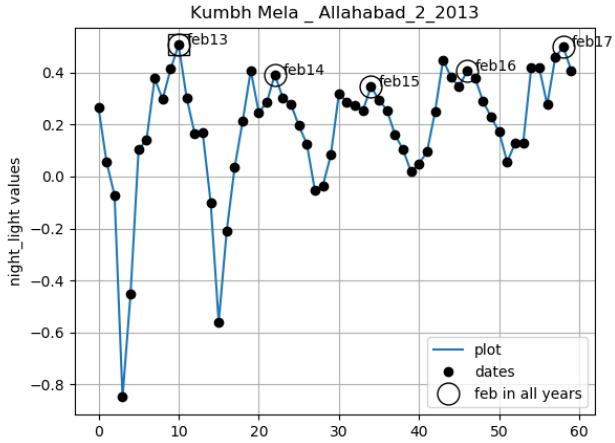

Fig. 5. Kumbh Mela in Allahabad in February 2013

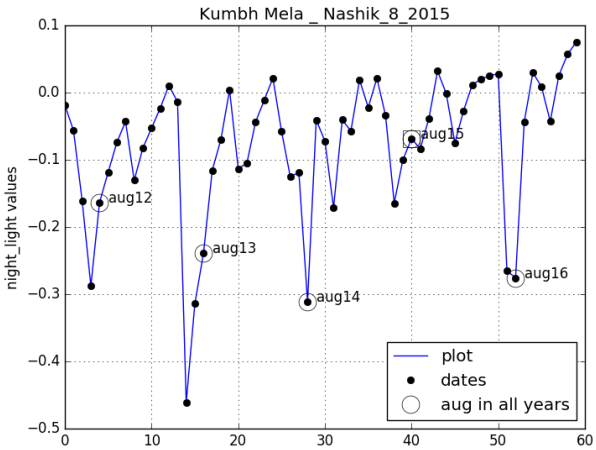

Fig. 6. Kumbh Mela in Nashik August 2015