Fuzzy Linear Regression for Tuberculosis Case Notification Rate Prediction in Surabaya

Siana Halim†
Industrial Engineering Department
Petra Christian University
Surabaya Indonesia
halim@petra.ac.id

Rolly Intan
Informatics Department
Petra Christian University
Surabaya Indonesia
rintan@petra.ac.id

Lily Puspa Dewi
Informatics Department
Petra Christian University
Surabaya Indonesia
lily@petra.ac.id

ABSTRACT

In this paper we discuss the fuzzy linear regression for predicting the Tuberculosis (TB) case notification rate prediction in Surabaya. The prediction of this disease is very important since TB is among the diseases that carries stigma because it is potentially highly contagious. In this research, we first describe the statistics for the case notification rate (CNR), particularly for the year 2017. We cluster the CNR for man, woman and total with respect to the poverty percentage in each district. Based on this clustering we can map in which district the Tuberculosis CNR and the poverty percentage is high, medium and low. Using the Moran I' statistics, we then tested the spatially dependency among the district. Based on the test result therefore, in this research we also regard the spatial effect of a district to the others in spreading the disease, and model the prediction using fuzzy linear regression. The model can predict good, the mean square error of the model is 835.87 and the mean absolute deviation is 22.46

CCS CONCEPTS

CCS → Computing methodologies → Artificial intelligence → Knowledge representation and reasoning → Vagueness and fuzzy logic

KEYWORDS

Fuzzy linear regression, local Moran I' Statistics, partitioning around medoids, tuberculosis, case notification rate

ACM Reference format:

Siana Halim, Rolly Intan and Lily Puspa Dewi. 2019. Fuzzy Linear Regression: for Tuberculosis Case Rate Notification in Surabaya. In *Proceedings of 2019 International Conference on Advanced Information Science and System (AISS'19). Singapore, 5 pages.*

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).
WOODSTOCK'18, June, 2018, El Paso, Texas USA
© 2018 Copyright held by the owner/author(s). 978-1-4503-0000-0/18/06...$15.00
https://doi.org/10.1145/1234567890
https://doi.org/10.1145/3373477.3373492

1 Introduction

Tuberculosis or TBC is one of the diseases that affects many people of Indonesia. A diagnosis of tuberculosis or TBC is among the most dreaded statements that anyone wants to hear. Apart from the likely risk of losing a promising job, TB is among the diseases that carries stigma because it is potentially highly contagious. Indonesia is listed as the world's third-largest TB-burdened country. East Java ranks second with several findings of 54,811 cases in 2017 [1].

The number of these case findings not only indicates that many patients in East Java, but the success of health workers to find tuberculosis cases, compared with the estimated number of patients in certain areas, which is familiarly called the Case Detection Rate (CDR). In East Java, the achievement of CDR in 2018 is 35,4% [2].

TB is an infectious disease caused by the bacillus Mycobacterium tuberculosis. It typically affects the lungs (pulmonary TB), but it can also affect other sites (extrapulmonary TB) such as brain, stomach, and skin. The disease is spread when people who are sick with pulmonary TBC expel bacteria into the air, for example by coughing. However, the risk factors for TBC spread also come from various environmental aspects such as low immune system, carrier in our environment, poor ventilation, poor nutrition, a place where many people gather (such as boarding schools or prisons) and densely populated villages. That is why big cities in East Java, such as Surabaya, has the highest number of TBC case.

Public education of TBC is increasingly urgent because of the attached stigma leading to self-denial, or our tendency to ignore "bad coughs". Public health experts fear many do not report related symptoms such as weight loss, shortness of breath, chills and other ailments. As a result, thousands of cases go unreported, and many more are categorized as multidrug resistant, while an effective vaccine for adults is yet to be developed. Poverty in the household means less priority on continuing treatment among TBC carriers as the focus is on short-term earnings even when one may feel unwell; while prosperous-looking cities may blind us to the fact that TBC can spread faster in dense, polluted areas. East Java Government

Department of Health initiate strategies to reduce the high TBC rate such as promotive, preventive and rehabilitative efforts by cooperating with the related partners in order to optimize the results, for example by providing information about the dangers of TB that are often not known to the public [3].

Many researches have been done in predicting the spread of the TBC, [4] predict the frequency and severity of the TBC meningitis immune reconstitution inflammatory syndrome. In this case they use logistic regression. Chen [5] proposes a model which links the disease progression, the related medical intervention actions and the logistics deployment altogether to support the decision-making process in case of the logistics response to an infectious disease from a strategic level. Ghosh [6] concluded from the analysis that the spread of the infectious disease increases when the growth of bacteria caused by conducive environmental discharge due to human sources increases. Cai [7] examine the use of various logistic growth curve models, via a series of simulated experiments in which the underlying true model is a mechanistic model of infectious disease spread.

In this case we use the fuzzy linear regression to predict the case notification rate in Surabaya [8]. We also include the spatial location as the predictor as well as the poverty percentage in the model. The fuzzy linear regression is robust for the case small sample [8] and in this research the data are collected from 63 community health centers in Surabaya therefore, the fuzzy linear regression is suitable to model the case.

2 Methods

In this study, we first did the data descriptive of the Surabaya population density, the area, percentage of poverty in each district (Kecamatan). There are 31 Kecamatan in Surabaya. Each Kecamatan consists of sub-district, so called Kelurahan. There are 154 Kelurahan in Surabaya. The data were collected in sixty-three community health centers (Puskesmas) in Surabaya. Each community health center cover one-two Kelurahans. We defined two community health centers as neighbor if they share a common boundary.

To test the spatial correlation of the TBC in each Puskesmas, we used the spatial local Moran I statistics. Anselin [9] suggested the local Moran's I statistics for identifying local clusters and local spatial outliers. The Local Moran's I statistics can be formulated as:

$$I_i = \frac{n(y_i-\bar{y})\sum_{j=1}^{n} w_{ij}(y_j-\bar{y})}{\sum_{i=1}^{n}(y_i-\bar{y})^2} \quad (1)$$

where y is the variable of interest. w_{ij} is the weight for the i, j observations.

2.1 Partitioning around Medoids (PAM)

Partitioning around medoids (PAM) is a clustering algorithm which is based on k-medoids instead of k-means. In the k-medoids clustering each cluster is represented by one of the data points in the cluster. These points are named as cluster medoids. The medoid is an object within a cluster in which average dissimilarity between the point and all other members of the cluster is minimal [10]. The k-medoids is a robust alternative to k-means clustering. The PAM algorithm [11], is the most common algorithm for performing the k-medoids clustering. Basically it consist of five steps: (1) Select k object to become a medoid, (2) calculate the dissimilarity matrix, (3) assign each object to its closest medoid, (4) for each cluster search if any object of the cluster can decrease the average dissimilarity coefficient, if it does, select the object, (5) if there is one medoid changed then go to (3) else end the algorithm.

2.2 Fuzzy Linear Model

In Fuzzy linear models, the dependent variable y and the predictors $x_i, i = 1, \dots m$ are fuzzy numbers, i.e., $\tilde{X}_i, \tilde{Y} \in \mathcal{R}_F, \forall i = 1, \dots, m$, where $\mathcal{R}_F$ is the space of fuzzy number. The fuzzy linear model can be written as [12]:

$$\tilde{Y} = \tilde{A}_0 \oplus (\tilde{A}_1 \otimes \tilde{X}_1) \oplus \dots \oplus (\tilde{A}_m \otimes \tilde{X}_m) \oplus \tilde{\delta} \quad (2)$$

where $\tilde{A}$ denotes fuzzy parameters of the regression model, $\tilde{A}_i \in \mathcal{R}_F, \forall i = 1, \dots, m$, $\tilde{\delta}$ is the error term, $\tilde{\delta} \in \mathcal{R}_F$. The detail models of the fuzzy linear models can be seen in [12].

The spatially dependence is included by measuring the distance between community health centers. Following the spatial autoregressive regression [13], then the disturbance of the first modeled is modeled as SAR, i.e.

$$\tilde{\delta} = \rho W \tilde{\delta} + \varepsilon \quad (3)$$

3 Results and Discussion

In general Surabaya with its 3.5 million populations (estimated) is the second largest city in Indonesia and the capital of East Java Province [14]. Surabaya has density reach 9900 peoples /km2 [15] (higher than Singapore 8108/km2 or Hong Kong 6677/km2). As a dense and a humid city (see Table 1), so TBC disease can spread easily in Surabaya.

Table 1: Surabaya Statistics In 2018

	Min	Mean	Max
Population (thousand)	12541	45802	87561
Area (Km2)	0.915	2.001	14.400
Density (thousand/Km2)	2733	46992	541022
Poverty percentage (%)	4.03	18.02	55.46
#Rainy day (days/month)	9.83	13.99	16.00
Precipitation (mm/month)	129.9	164.6	194.9
Max Humidity per month	70	88.72	94.75
Min Humidity per month	46.08	53.14	57.83
Max Temperature	28.21	33.30	34.43
Min Temperature	23.11	26.29	28.73

*Summaries form [16]

The data were collected from 63 community health centers in Surabaya from 2012 up to 2017. The recorded data include population (number of men and the number of women) of each district in which the community health located, the area, the poverty percentage, number of TB infected, men and women, and the case notification rate (CNR) for men and women in each region per 100000 population. The CNR is calculated as number of TB infected divide by total population for each gender times 100.000, i.e., [17]

$$CNRMan2017 = \frac{\#TBinfected\ Man}{\#Man\ Population} \text{ x } 100000 \quad (4)$$

$$CNRWoman2017 = \frac{\#TBinfected\ Woman}{\#Woman\ Population} \text{ x } 100000 \quad (5)$$

$$CNRTotal2017 = \frac{\#TBinfected}{\#Population} \text{ x } 10000 \quad (6)$$

It can be seen in Fig.1 the case notification rate for man is higher than for woman. We then use the partitioning around medoids (PAM) to cluster the poverty percentage and CNRMAN 2017 as well as CNRWOMAN 2017. Based on the minimum weighted sum square we got that the number of clusters is three. Let suppose we named the cluster as Low, Middle and High. The mean of the poverty percentage and the CNR of each cluster is presented in the Table 2.

Table 2: Mean of the poverty percentage and CNR

Cluster	Man		Woman		Total	
	Poverty	CNR	Poverty	CNR	Poverty	CNR
Low	13.22	79.77	12.33	64.12	12.47	74.58
Middle	18.19	156.20	17.94	112.34	15.60	150.05
High	23.02	224.10	27.44	162.67	25.04	214.96

The CNR maps of those clustering are depicted in Figure 1, 2 and 3. We can see that there are 21 regions categorize as high case notification rate for man. The top five regions are in Tambak Wedi (T1), Bulak Banteng (B6), Balongsari (B2), Simolawang (S7) and Morokrembangan (M6). Looking at the map, we know that those regions are in the north of Surabaya. The north of Surabaya is a water catchment area, old villages and maritime area. It is bordered to the Java Sea and many fishermen lived in those area. The poverty percentage on those areas are higher compare to the other areas in Surabaya. While for woman case, there are 16 regions categorize as high CNR. The top five regions are in Sawah Pulo (S1), Banyu Urip (B4), Putat Jaya (P7), Jeruk (J3) and Dupak (D3). In total (both man and woman CNR) there are 23 regions categorize as high CNR. The top five regions with highest CNR are in Tambak Wedi (T1), Bulak Banteng (B6), Balongsari (B2), Simolawang (S7) and Morokrembangan (M6). Moreover, the local Moran'I statistics shows that there is spatial correlation among the regions. The

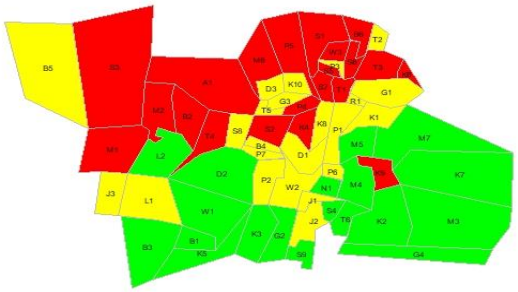

Figure 1: The map of CNR Man 2017

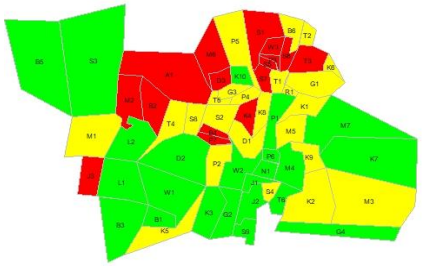

Figure 2: The map of CNR Woman 2017

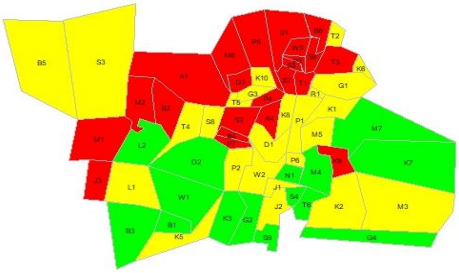

Figure 3: The map of CNR Total 2017

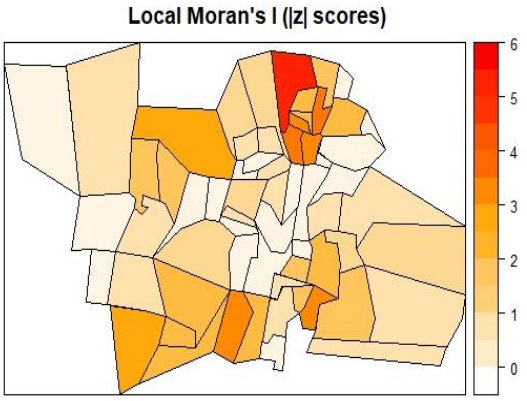

Figure 4: Local Moran's I score

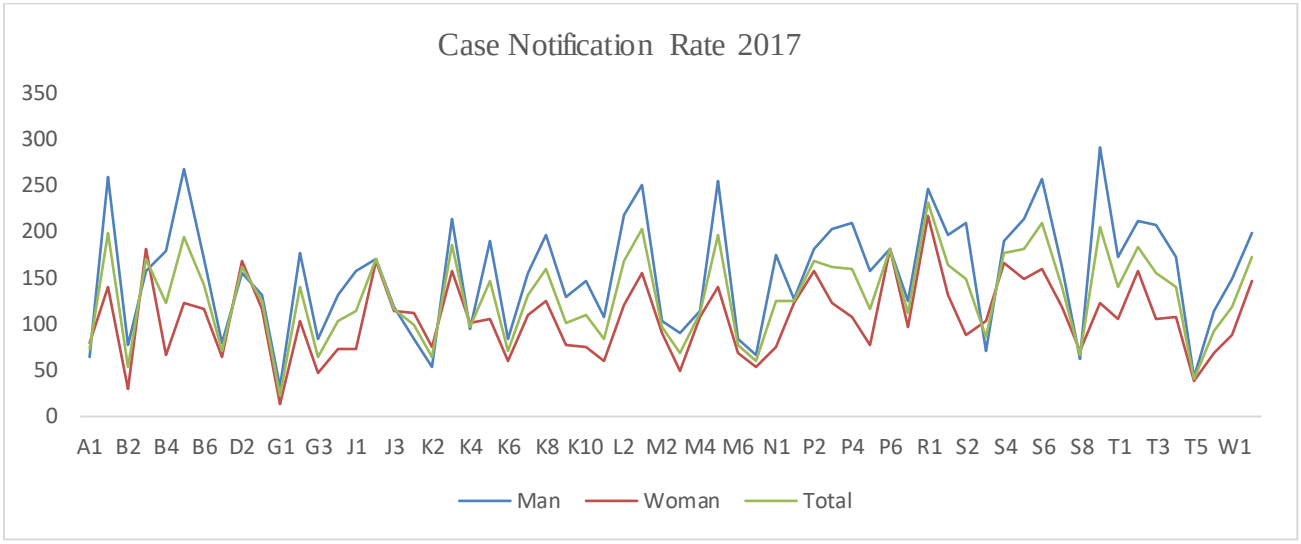

Figure 5: TBC Case Notification Rate 2017 for in Surabaya's Puskesmas

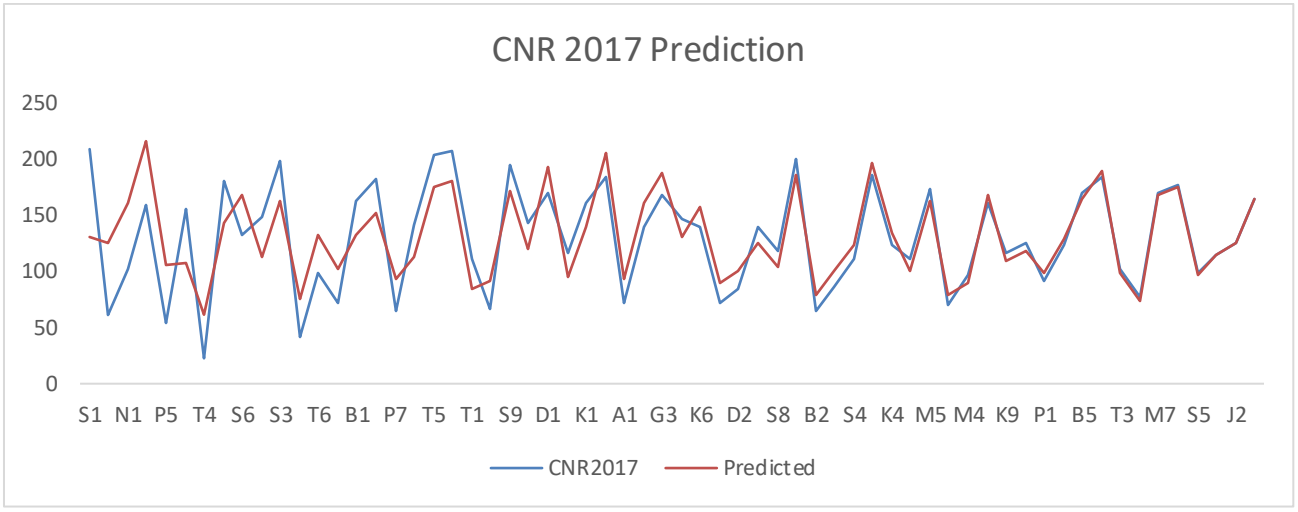

Figure 6: The CNR 2017 prediction

degree of correlation is varied from each location and it is shown in Fig. 4

It is well known that the fuzzy regression provides an alternative to the regression when the model is indefinite, the relationships between model parameters are vague, sample size is low or when the data are hierarchically structured [8]. Therefore, in this case the fuzzy regression linear model is suitable, since the data set is small only 63 with many predictor variables as presented in Fig. 5.

Finally, we modeled the CNRTotal 2017 prediction using the fuzzy linear model. It is well known that the fuzzy regression provides an alternative to the regression when the model is indefinite, the relationships between model parameters are vague, sample size is low or when the data are hierarchically structured [8]. Therefore, in this case the fuzzy regression linear model is suitable, since the data set is small only 63 with many predictor variables as presented in Fig. 5. The CNRTotal 2017 is predicted via CNRTotal 2012 to CNRTotal 2016 and poverty percentage. The spatially dependencies in this data is measured using the weighted matrix. If a region is closed to each other (sharing boundary) then the weight is one, otherwise it is defined as zero.

The result of the prediction and the CNRTotal 2017 is depicted in Fig 6. For several regions the prediction values are closed to the actual ones. However, for some regions such as Sawah Pulo, Simolawang, Ngagelrejo, Krembangan Selatan, Perak Timur and Bangkingan. In general, the mean square error of the model is 835.87 and the mean absolute deviation is 22.46.

This research is not yet finished. In this paper we only model using the fuzzy linear model and concern with the spatially

dependency in the error term is measured as 0-1 (crisp). In the future research we will consider the distance between two neighborhoods in a fuzzy approach.

4 Conclusion

This paper presented a model to predict case notification rate for the Tuberculosis spreading in Surabaya. The fuzzy linear model is chosen since the data set is only 63, which are came from the Surabaya community health centers (Puskesmas) in 6 years (2012-2017). The result shows that the fuzzy linear regression can predict the CNR closed enough to the actual values for several regions, and not closed enough to some other regions. The mean square error of this model is 835.87 and the mad is 22.46. In the future work, we would like to extend the fuzziness in the neighborhood distance measurement.

ACKNOWLEDGEMENT

The authors are very grateful to the Directorate General of Higher Education of the Republic of Indonesia for supporting this research.

REFERENCES

[1] Dinas Kesehatan Provinsi Jawa Timur (2018). Profil Kesehatan Provinsi Jawa Timur tahun 2017, Kementrian Kesehatan Republik Indonesia, Surabaya, Indonesia.

[2] S. Anung (2018). Evaluasi rencana aksi 2018 TBC, imunisasi dan rencana tindaklanjut Tahun 2019, di Pertemuan RAKORPOP Kementrian Kesehatan, 23 November 2018.

[3] Dinas Kesehatan Surabaya, Dinas kesehatan Surabaya didorong memaksimalkan penanganan TBC, https://www.antaranews.com/berita/697537/dinkes-surabaya-didorong-maksimalkan-penanganan-tbc.

[4] S. Marais, G. Meintjes, D.J. Pepper, L.E. Dodd, C. Schutz, Z. Ismail, K.A., Wilkinson and R.J. Wilkinson (2013). Frequency Severity and Prediction of Tuberculous Meningitis Immune Reconstitution Inflammatory Syndrome. Clinical Infectious Diseases, 56(3), 450-460.

[5] W. Chena, A. Gurnetb and A. Ruizc (2015). Modelling the Logistics Response to a General Infectious Disease, IFAC-Papers OnLine, 48(3), 180-186.

[6] M. Ghosha, P Chandraa, P. Sinhaa and J.B. Shuklab (2006). Modelling the Spread of Bacterial Infectious Disease with Environmental Effect in a Logistically Growing Human Population. Nonlinear Analysis: Real World Applications, 7, 341 – 363.

[7] L. Cai, 2013. Logistic Growth Models for Estimating Vaccination Effects in Infectious Disease Transmission Experiments. Thesis, The University of Guelph, Guelph, Ontario, Canada.

[8] P. Skrab´anek and N. Mart´ınkov´a, (2018). Fuzzyreg: An R Package for ˇ Fuzzy Linear Regression. In: Cech P., Svozil D. (eds.), ˇ ENBIK2018 Conference Proceedings, Prague, 7.

[9] L. Anselin (1995), Local Indicators of Spatial Association-LISA. Geographical Analysis, 27, 93-115.

[10] A. Kassambara (2017) Practical Guide to Cluster Analysis in R: Unsupervised Machine Learning, STHDA, http://www.sthda.com

[11] L. Kaufman and P.J. Rousseeuw (1990), Finding group in data: An introduction to cluster analysis, Wiley, New York.

[12] P. Skrab´anek, J. Marek (2018), Model used in fuzzy linear regression APLIMAT2018 Conference Proceedings, Slovak University of Technology in Bratislava

[13] L. Anselin (1988), Spatial econometrics: methods and models. Kluwer Academic Publishers, Dordrecht, The Netherlands

[14] Statistics Indonesia, (2017) Surabaya Municipality in Figures 2017, surabayakota.bps.go.id.

[15] Indonesia: Java (Regencies, Cities and Districts) – Population Statistics, Charts and Map.

[16] BPS, Surabaya dalam Angka. Biro Pusat Statistika Surabaya, https://surabayakota.bps.go.id/publication/2018/08/21/35de76f19338e3ecd445b838/kota-surabaya-dalam-angka-2018.html

[17] Depkes RI, Profil kesehatan Surabaya tahun 2017, http://www.depkes.go.id/resources/download/profil/