Detection of Towns Having a Peculiarity by Using Regression Models

Dr. Noriaki Sakamoto
Hosei University, Japan

Doi:10.19044/esj.2019.v15n10p113 URL:http://dx.doi.org/10.19044/esj.2019.v15n10p113

Abstract

This paper proposes a method to detect towns having a peculiarity, which is a statistical outlier from a statistical table. A statistic often contains data that are peculiar and are also known as outliers which are followed as large residuals in regression models. The detection of outliers in statistical tables was studied. The table has 22 explanatory variables, one response variable and 1947 records which can clarify their efficient causes or mixed effects. This information have greatly helped local governments with their policy and improvement of each region, for example; infrastructures, public services, and subsidies or grants. Although many studies have been made on grouping records or building a predictive model to overcome outliers, little attention has been given to find outliers. Many of those studies require a model's parameter tuning and learning, or a description of a fitting function. Furthermore, for municipal officers to find outliers, it would be desirable to be able to analyze readily Free Software R without programming. Therefore, we propose a method to detect outlier from a statistical table by using three regression models which do not require learning and parameter adjustment provided by R.

Keywords: Outlier, Statistics, Regression model, Additive regression model, Robust regression model, Data mining

Introduction

This paper proposes a method to find towns having a peculiarity in a statistical table. First, Table 1 shows the contents of the statistical table to be studied in this paper.

Table 1. Variable Names of a statistical table used for this paper

	Town No.	Town Address	Explanatory Variable No.1	---	Explanatory Variable No.22	Response Variable (No.23)
Record* →	1	Address	Value	- - -	- - -	- - -
	2	- - -	- - -	- - -	- - -	- - -
	- - -	- - -	- - -	- - -	- - -	- - -
	1947	- - -	- - -	- - -	- - -	- - -

*Record number is equal to "Town No".

The response variable is an Inflow Potential Index with population movement defined by Mori (2018). We abbreviated the index to IPI, and IPI is calculated by Equation (1) for each town.

$$\mathrm{IPI} \equiv {M_{si}} \Big/ {\sum_g \left(P_{sgi} R_{sg}\right)} \qquad \textbf{(1)}$$

s: Sex

g: Age

i: Town i area in the ward

M: The number of inflow residents per Town i area

P: Population

R: City average of M

The trend for population movement is not only sex and age but is also as a result of various efficient causes. These causes are the aggregate result in the real population movement in each town. In other words, the movements into town differ because of various efficient causes even if the population of the town has similar sex and age composition. Explanatory variables in the statistical table (Table 1, data available from "The 2015 Population Census of Japan") show the efficient causes (Mori, 2018). Detecting outliers in this statistical table makes it possible to clarify effective causes and/or mixed effects other than sex and age. This information greatly helps local governments with their policy and improvement of each region, for example; infrastructures, public services, and subsidies or grants (Kojima, 2013; Koike, 2018).

However, previous studies have focused on Analysis, reduction of explanatory variables or Predictive modeling, and few studies on detection.

Analysis

For example, Cluster analysis classifies records (Christopher Chatfield & Alexander Collins, 1980a; Williams Allan, M., *et al*., 2017). Thus, by looking at the dendrogram of the cluster analysis, which is the result of classification, we might be able to find outlying towns (i.e. statistical outliers). However, Cluster analysis has various choices that include "data normalization", analysis methods such as "single chain method, group average method, word method, minimum variance method", and selection of distances and so on. Therefore, since we obtain different results based on that choice, it is difficult to judge that detection is successful. In addition, we cannot draw a dendrogram of Cluster Analysis of 1947 records of the statistical table.

Reduction of Explanatory Variables

Since there are many explanatory variables in Table 1, we usually use Principal Component Analysis to reduce the number of variables (Christopher Chatfield & Alexander Collins, 1980b; Salvador García., *et al*., 2014). However, the reduction of explanatory variables is not used for the following reasons in this paper:

1) The causal relationship of explanatory variables is unclear.
2) The reduced explanatory variable may affect the objective variable.

Predictive Modeling

It is possible to obtain a prediction model by using machine learning (Svein Nordbotten, 1996; Matthew Sadiku, N. O., *et al*., 2015; Bruce Ratner, 2017). For example, the following citations are part of the survey paper (Hossein Hassani & Emmanuel Sirimal Silva, 2015) which stated that "an imputation based on Neural Network model was applied to the Norwegian population census data of 1990 in order to perform a population census by combining administrative data along with data gathered through sample surveys." "Cluster Analysis was used as a method for predicting missing data by analyzing the 2007 census donor pool screening." Other research, for example by Sawada (2016), constructed a model to calculate an estimated regional population by using Support Vector Machine. However, these approaches give the following problems to the purpose of our study.

1) The purpose of many studies of applying machine learning to statistical data is to overcome outliers such as population estimates, economic indicators, and predictions of missing data, etc.

2) Machine learning requires parameter adjustment, and it is necessary to repeat a simulation for learning.

3) All models or parameters that succeeded in learning are not identical.

Detecting Outliers

Harvey Motulsky and Ronald Brown (2006) propose a method for identifying outliers which combines robust regression and outlier removal. This is based on the assumption that scatters following Lorentzian distribution or Gaussian distribution. Ogu, A. I., *et al*. (2013) study detects outliers in a univariate and bivariate data. Previous studies also have these limitations. The distribution function of statistical data we deal with is unknown, and the explanatory variable is 22.

Therefore, as a result of applying three regression models that do not require learning and parameter adjustment, we propose a method of finding outlier records from a statistical table by considering records with large residuals as outlier records. Figure 1 explains our idea. The data that is (circled part) away from the approximate straight line is an outlier, while the residual between the data and the predictive model drawn by the straight line is large. In other words, data with large residual is an outlier and we consider that it has a peculiarity. The three regression models are multiple regression, additive regression and robust regression provided by Free Software R (The Comprehensive R Archive Network; An Introduction to R; The R Project for Statistical Computing (Hadley Wickham & Garrett Grolemund, 2017). The results of applying three regression models to a real statistical table show the effectiveness of our method.

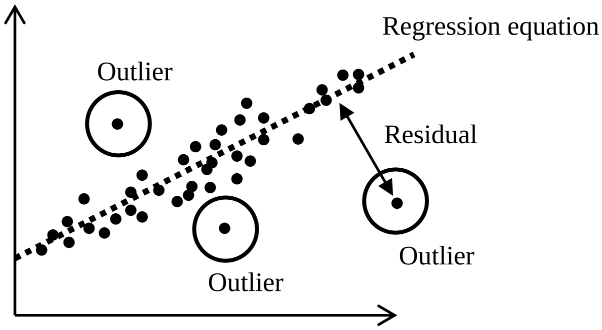

Figure 1. Outlying Data with a large residual of a regression equation.

Statistical Table

In order to examine various effective causes other than sex and age, Mori (2018) acquired explanatory variables and data shown in Table 2 and 3 from the data of each town in Niigata City, Niigata Prefecture, and “The 2015 Population Census of JAPAN”. Thus, the number of records is 1947 towns. On the other hand, IPI which is a response variable is calculated from Equation (1) for each town.

Explanatory variables and Response variable have no causal relationship and they are independent (See Equation (1)). Response variables are calculated without using explanatory variables. Therefore, finding a mathematical model is difficult.

For example, when using a generalized linear model, an approximate function (or a fitting function) such as Gaussian function, Poisson function, and binomial function must be explicitly specified. Therefore, we apply three regression models (Linear model: lm, Generalized Additive Model: gam, Robust Fitting of Linear Model: rlm) without using a regression model (Mixed model, Local approximate regression, etc.) that needs the description of a fitting function. To calculate three regression models, we use R.

Table 2. Explanatory variables

No.	Content: Rate of	No.	Content: Rate of	
1	Never Married	12	Manufacturing workers	
2	Married	13	Service workers (A)[1)]	
3	One-person households	14	Public Employees	
4	Householder	15	Administrative and managerial Workers	
5	Rented house	16	Service workers (B)[2)]	
6	Detached house	17	Agricultural, forestry, and fishery workers	
7	Apartment house or flat	18	Manufacturing process workers	
8	Three-story or higher house	19	The period of living in the current house:	Less than 1 years
9	Employee	20		Less than 5 years
10	Self-employed worker and Family worker	21		5 to 20 years old
11	Agriculture and Forestry workers	22		20 years old or more

1) Service workers
- Scientific research, professional, and technical services
- Accommodations, eating, and drinking services
- Living-related and personal services and amusement services
- Education, learning support
- Medical, health care, and welfare
- Compound services

2) Service workers
- Home life support services
- Nursing-care services
- Health care services
- Life health services
- Customer service
- Building custodial service

Table 3. Some of the 1947 records

Town No.	Town Address[*)]	Explanatory variables				Response variable
		No.1	No.4	No.13	No.22	No.23 : IPI
1	Tarodai, Kita Ward	0.2422	0.9421	0.3345	0.5087	0.2714
- - -	- - -	- - -	- - -	- - -	- - -	- - -
500	Gakkoura Town, Chuo Ward	0.3421	0.7000	0.2500	0.0251	2.3781
- - -	- - -	- - -	- - -	- - -	- - -	- - -
1000	Satsuki Town 2, Konan Ward	0.2554	0.7869	0.4182	0.5343	0.6564
- - -	- - -	- - -	- - -	- - -	- - -	- - -
1947	Warimae, Nishikan Ward	0.2000	0.9730	0.4167	0.4870	1.0182

*) Niigata City, Niigata Prefecture, Japan

Detection of Towns having a Peculiarity

The process of detection is as follows:

Step1: Regression analysis.

Step 2: Calculate the residual for each record (See Figure 2).

Residual = IPI (*Response variable*) - *prediction value by regression equation* (*Calculate it by Explanatory variables*) (2)

Step 3: Detect the larger residual top 10 towns.

Table 4 shows the three regression results (lm, gam, rlm) calculated by R and that there is a very strong correlation. The existence of a strong correlation does not imply a causal link between the variables. However, our research objective is not to obtain a causal model or a mathematical model representing causality.

Table 5 shows the 10 detected towns. In the data in Table 5, the town common to the results of the three regression models is shown in bold. The towns of the bold type have a singularity.

Table 4. Three regression results

lm	gam	rlm
Residual standard error: 0.1679 Multiple R-squared: 0.9076 Adjusted R-squared: 0.9065	Adjusted R-squared: 0.95 Deviance explained: 95.3% GCV: 0.015926	Residual standard error: 0.08616

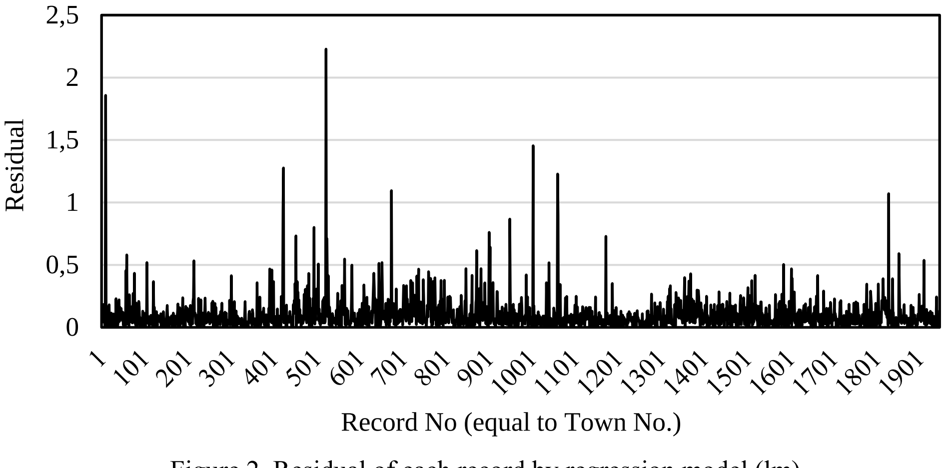

Figure 2. Residual of each record by regression model (lm)

Table 5. The larger residual top 10 towns

lm		gam		Rlm	
Residual	Town No.	Residual	Town No.	Residual	Town No.
2.23	**522**	1.36	**1060**	2.25	**522**
1.86	**10**	1.33	**522**	1.94	**10**
1.45	1003	0.93	**674**	1.64	1003
1.28	423	0.88	723	1.50	**1829**
1.23	**1060**	0.70	**10**	1.37	423
1.09	**674**	0.67	524	1.26	**1060**
1.07	**1829**	0.64	**1829**	1.12	422
0.94	422	0.63	452	1.08	1061
0.90	1061	0.56	1853	1.07	**674**
0.87	949	0.56	920	1.04	949

Town No: Address

10: Hamamatsu Town, Kita Ward, Niigata City, Niigata Prefecture, Japan
522: Jindouji 2, Chuo Ward, Niigata City, Niigata Prefecture, Japan
674: Nishiborimaedori 9 Town, Chuo Ward, Niigata City, Niigata Prefecture, Japan
1060: Hayadori 6, Konan Ward, Niigata City, Niigata Prefecture, Japan
1829: Takeno Town 1, Nishikanku Ward, Niigata City, Niigata Prefecture, Japan

Discussion of Detected Towns

First, Figure 3 and 4 shows the top 20 towns and the least 20 towns of IPI respectively. This is done in order to know the specificity of high-value towns and low-value towns of IPI. The two figures clarify the following,

Peculiarities of Towns with High IPI (See Figure 3)

High1: No. 1 to No. 9 draws sawtooth wave graphs. It refers to a married person, a detached house owner, indicating that the employment rate is high.

High2: If the values of No. 20 (Less than 5 years) is less than the values of No. 21 (5 to 20 years old), then IPI becomes large. Many residents have less than 5 years.

Peculiarities of Towns with Low IPI (See Figure 4):

Low: If the values of No.20 (Less than 5 years) is greater than the values of No.21 (5 to 20 years old), then IPI becomes small. Many residents are more than 5 years.

Next, we consider the detected towns. To find the peculiarities, the graphs in Figures 5 to 14 compare the adjacent to towns and towns of the same population. Table 6 shows the result of that consideration. Furthermore, we consider it from the actual town by looking at "Google Map," and we added the result to Table 6.

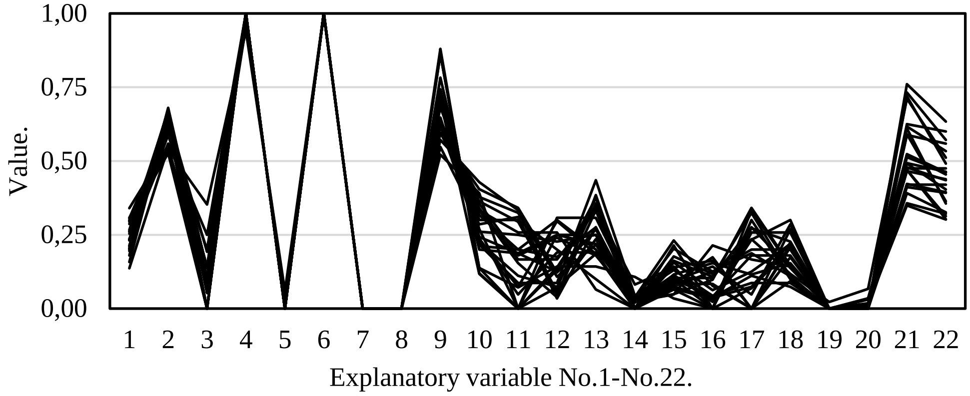

Figure 3. The top 20 towns of IPI

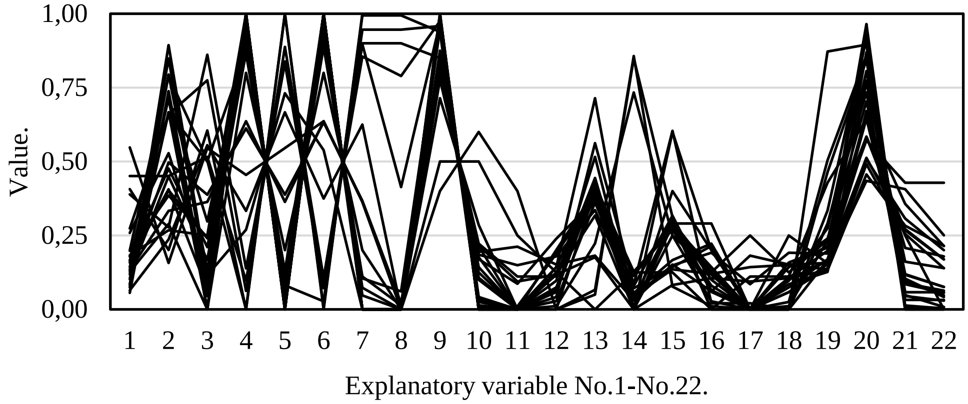

Figure 4. The bottom 20 towns of IPI

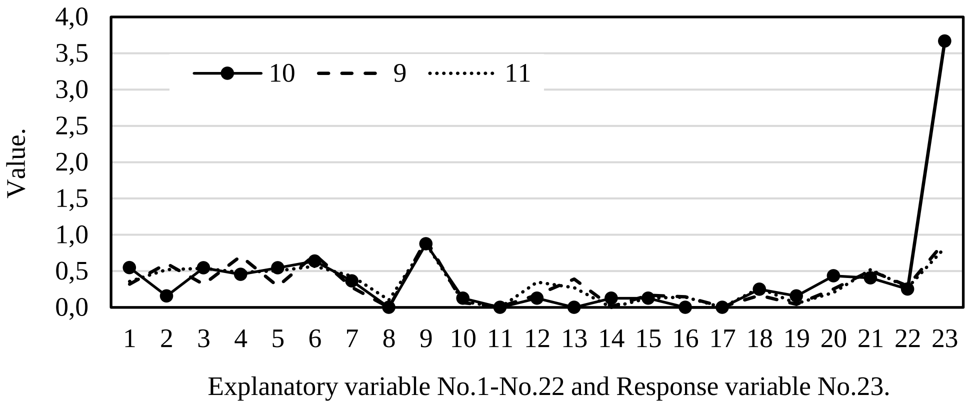

Figure 5. Town No.10 detected, Town No.9, and No.11 adjacent to that town

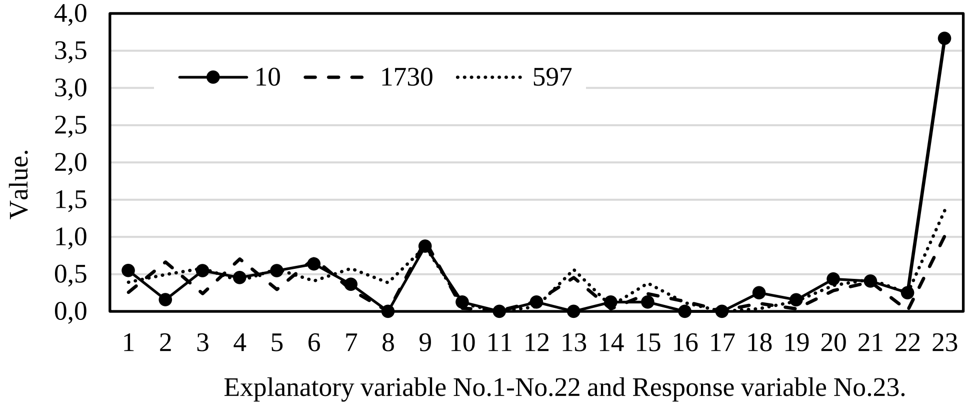

Figure 6. Town No.10 detected, Town No.1730, and No.597 of the same population.

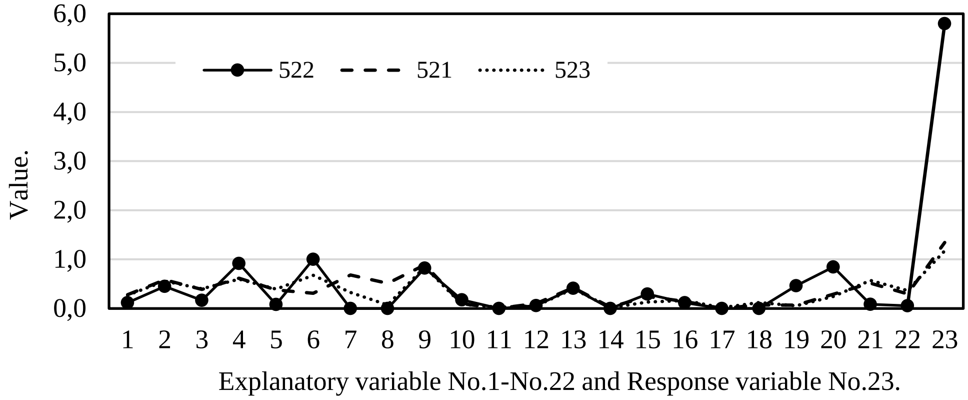

Figure 7. Town No.522 detected, Town No.521, and No.523 adjacent to that town.

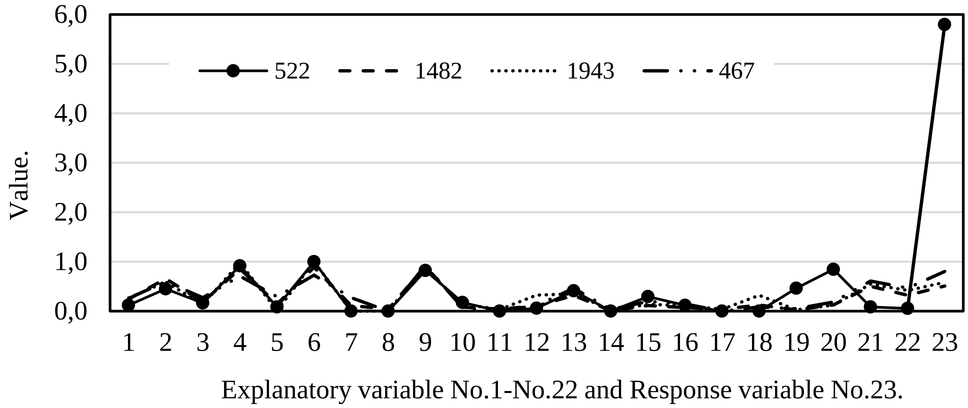

Figure 8. Town No.522 detected, Town No.1482, No.1943, and No.467 of the same population as that town.

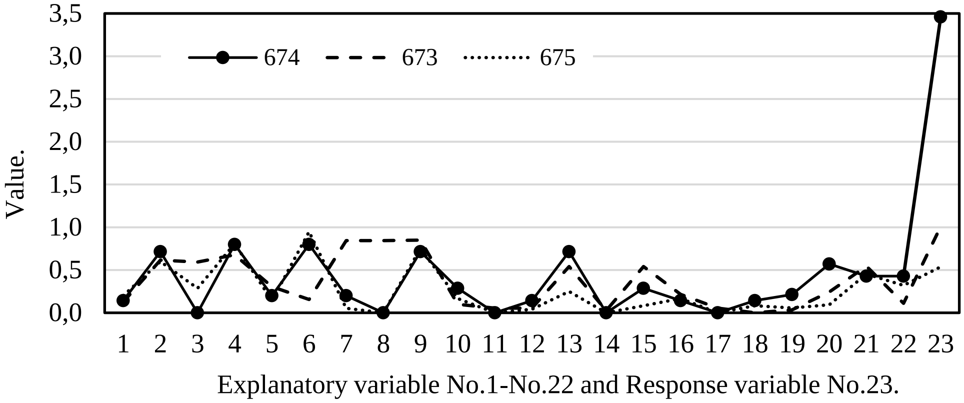

Figure 9. Town No.674 detected, Town No.673, and No.675 adjacent to that town.

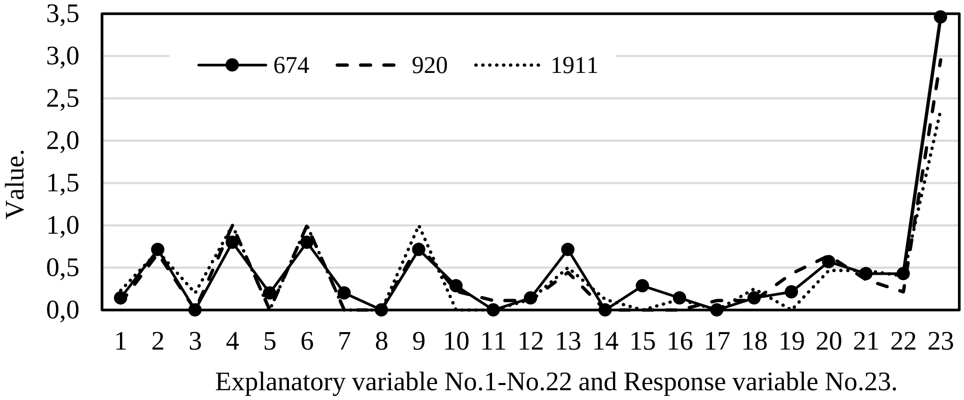

Figure 10. Town No.674 detected, Town No.920, and No.1911 of the same population as that town.

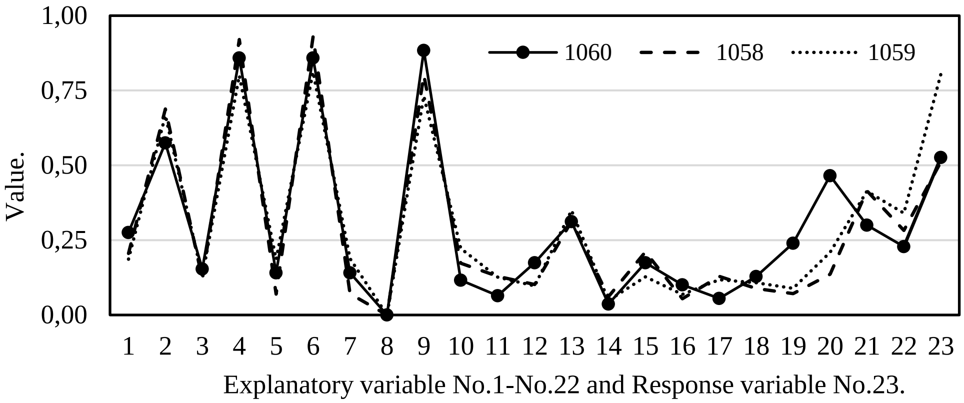

Figure 11. Town No.1060 detected, Town No.1058, and No.1059 adjacent to that town.

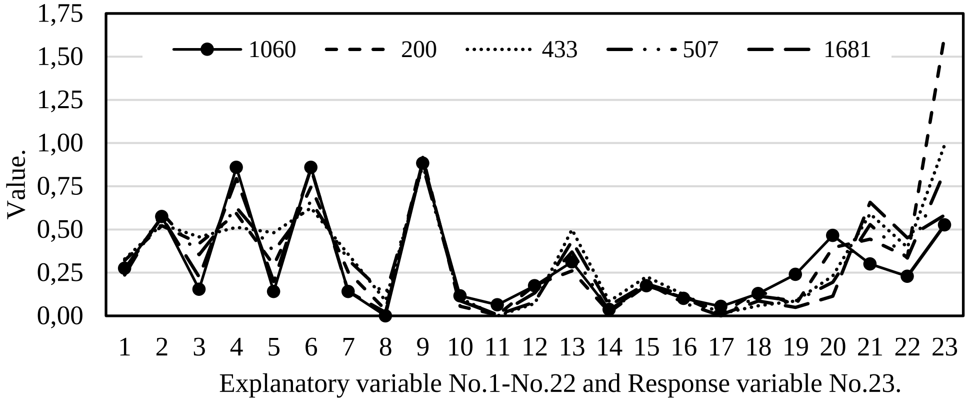

Figure 12. Town No.1060 detected, Town No.200, No.433, No.507, and No.1681 of the same population as that town.

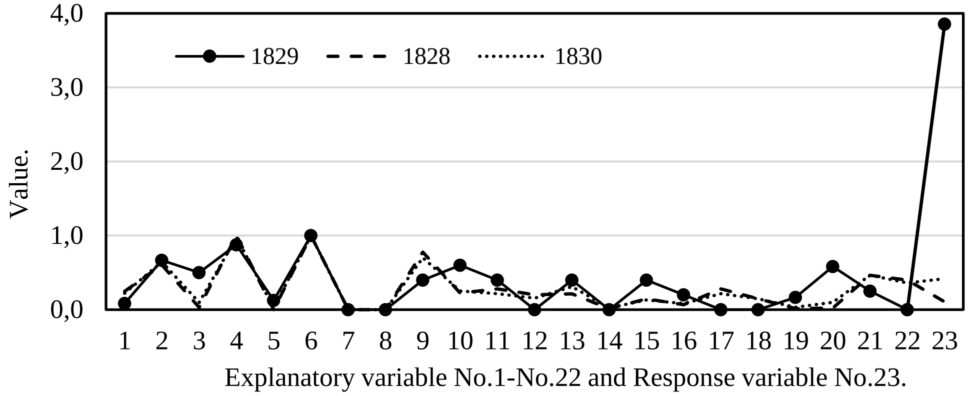

Figure 13. Town No.1829 detected, Town No.1828, and No.1830 adjacent to that town.

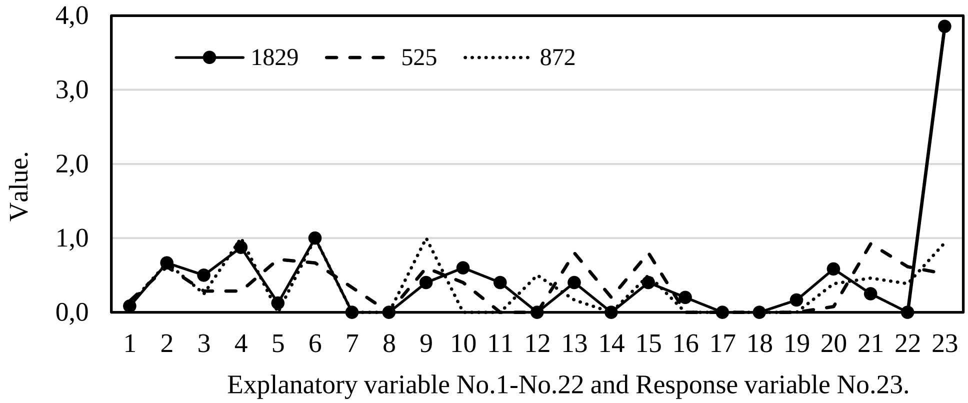

Figure 14. Town No. 1829 detected, Town No. 525, and No. 872 of the same population as that town.

Table 6. Consideration that the detected town is different from other towns

Town No and Address	Figure No	Compared to another adjacent town. Compared to other towns of the same population. / the specificity by looking at the Google Map
Town No. 10 Hamamatsu Town, Kita Ward	No. 5 No. 6	Although the values of No. 1 to No. 22 are almost equal, IPI is very large. The trend of No. 1 and No. 2 are reversed.
		There is a big factory, and the company dormitory for singles. We consider that specificity is the influence of the workers living there. As a corroboration, the values of No. 1 and No. 18 are large.
Town No. 522 Jindouji 2, Chuo War d	No. 7 No. 8	Town No. 522 is the maximum value of IPI. Since the values of No. 20 is greater than the values of No. 21, then IPI become small (See *Peculiarities of towns with low IPI*: **Low**), but IPI is very large.
		We consider that specificity is the influence of the new residential area, and the nursing home (long-term care health facility) living in a short period of years.
Town No. 674 Nishiborimae-dori 9 Town, Chuo Ward	No. 9 No.10	Although the values of No. 1 to No. 22 are almost equal, IPI is large. The value of No.13 (Service workers (A)), and No. 15 (Administrative and managerial workers) are large.
		There are many large Japanese-style restaurants. We consider that specificity is the influence of the workers living there. As a corroboration, the values of No. 2 and No. 13 are large.
Town No.1060 Hayadori 6, Konan Ward	No.11 No.12	Since the values of No. 20 is greater than the values of No. 21, then IPI become small (See *Peculiarities of towns with low IPI:* **Low**), but IPI is large.
		We consider that specificity is the influence of the nursing home (long-term care health facility) living in a short period of years.
Town No.1829 Takeno Town 1 Nishikanku Ward	No.13 No.14	Although the values of No. 1 to No. 22 are almost equal, IPI is very large. Since the values of No. 20 is greater than the values of No. 21, then IPI become small (See *Peculiarities of towns with low IPI*: **Low**), but IPI is very large.
		We consider that specificity is the influence of the new residential area.

Conclusion

This paper proposes a method of finding outlier records from a statistical table by considering records with large residuals as outlier records by using a regression model that does not require learning and parameter adjusting. The approach used in this paper does not require programming for mathematical calculations and can be easily implemented using Free Software R. The proposed method is applied to the statistical table on the movement of the population, and the knowledge about its specificity is obtained. Therefore, we have identified the company's single dormitory, the new residential area, and the presence of a nursing home as a feature of the outlying town.

References:

1. An Introduction to R. https://cran.r-project.org/doc/manuals/r-release/R-intro.html#Some-non_002dstandard-models, 11.2 Linear models, 11.8 Some non-standard models. (accessed 20 February 2019).
2. Bruce Ratner (2017). Statistical and Machine-Learning Data Mining: Techniques for Better Predictive Modeling and Analysis of Big Data, 3rd Edition, CRC Press.
3. Christopher Chatfield & Alexander Collins, J. (1980a). Cluster analysis, Introduction to Multivariate Analysis. New York: Routledge, Chapter 11, https://doi.org/10.1201/9780203749999.
4. Christopher Chatfield & Alexander Collins, J. (1980b). Cluster analysis, Introduction to Multivariate Analysis. New York: Routledge, Principal component analysis, Chapter 4 (principal component analysis), Chapter 8 (factor analysis), https://doi.org/10.1201/9780203749999.
5. Google Map. https://www.google.com/maps/. (accessed 20 February 2019).
6. Hossein Hassani & Emmanuel Sirimal Silva (2015). Forecasting with Big Data: A Review, Springer-Verlag Berlin Heidelberg, Sec.4.3, https://doi.org/10.1007/s40745-015-0029-9.
7. Harvey Motulsky, J. & Ronald Brown, E. (2006). Detecting outliers when fitting data with nonlinear regression-a new method based on robust nonlinear regression and the false discovery rate, BMC Bioinformatics 7:123, https://doi.org/10.1186/1471-2105-7-123.
8. Hadley Wickham & Garrett Grolemund (2017). R for Data Science: Import, Tidy, Transform, Visualize, and Model Data, O'Reilly Media, Chapter.23 Model basics.
9. Kojima, K. (2013). Analysis of Migrated Elderly Living in Private Households - Analysis using the Micro-data of “The 7th National Survey on Migration (2011)”, Journal of Population Problems, National Institute of Population and Social Security Research, Vol.69, No.4, pp.25-43. (in Japanese).
10. Koike, S. (2018). Demographical Migration Analysis of 20 Cities in Niigata Prefecture: Part 1: Overview of the Migration Pattern Change from 1980 to 2015, Journal of Population Problems, National Institute of Population and Social Security Research, Vol.74, No.1, pp.42-60. (in Japanese).
11. Mori, H. (2018). Detection of outlying regional units by Interregional Inflow Migration Potential Index, Japan Statistics Research Institute at Hosei University, Occasional Papers, No.94. (in Japanese).

12. Matthew Sadiku, N. O., Adebowale Shadare, E. & Sarhan Musa, M. (2015). DATA MINING: A BRIEF INTRODUCTION, European Scientific Journal, Vol.11, No.21, pp.509-513.
13. Ogu, A. I., Inyama, S.C., & Achugamonu, P.C. (2013). Methods of Detecting Outliers in A Regression Analysis Model, West African Journal of Industrial and Academic Research, Vol.7, No.1, pp.105-112.
14. Salvador García, Julián Luengo & Francisco Herrera (2014). Data Reduction, Data Preprocessing in Data Mining, Springer, Chapter 6, pp.147-162.
15. Svein Nordbotten (1996). Neural Network Imputation Applied to the Norwegian 1990 Population Census Data, Journal of Official Statistics, Vol.12, No.4, pp.385-401.
16. Sawada, T. (2016). Estimation of a regional population using Support Vector Regression, The Bulletin of the Centers for Information Media Studies, Aichi University, Vol.26, No.1, pp.31-47. (in Japanese).
17. The 2015 Population Census of JAPAN, https://www.stat.go.jp/english/data/kokusei/index.html. (accessed 20 February 2019).
18. The Comprehensive R Archive Network, https://cran.r-project.org/. (accessed 20 February 2019).
19. The R Project for Statistical Computing. https://www.r-project.org/, (accessed 20 February 2019).
20. Williams Allan, M., Jephcote Calvin., Janta Hanja & Li Gang (2017). The migration intentions of young adults in Europe: A comparative, multilevel analysis. Population Space Place., https://doi.org/10.1002/psp.2123.