IJITEB
Vol. 5
No. 1
2023

International Journal of Information Technology and Business

http://ejournal.uksw.edu/ijiteb

Analysis of Data Aggregates of Social Demography

Atmoko Nugroho[a]*

[a] Informatic Engineering, Faculty of Information and Communication Technology, Universitas Semarang, Indonesia

Abstract : Social demographics in an area are often displayed in the form of aggregate data for all those who want social demographic data. But often the components displayed are not uniform between one region and another, after all, the relationship between the components is not explained. Therefore, the analysis is needed to determine the interrelationships between the components and to test the aggregate data that is displayed.

Keywords: social demography, aggregates, interrelationship, region

1. Introduction

The website is a part of socialization that is done or requested by the government for its citizens. That way citizens who want information can get it on the government website. Information that is often sought by citizens is population and social information. And population and social information may not all be displayed, given some information that can have adverse effects. The Indonesian government also applies this, so that not all data or information displayed is raised. Besides, if the overall data displayed on a website is also ineffective or inefficient, this is because the population of Indonesia is large, diverse and has a complex geographical location. Therefore, the government website only displays aggregate data, it can be in the form of sample data, average data or total data [3]. One website with another website in one department may be different components displayed, depending on the policy decision-makers of the website manager [7].

The demography comes from Greek which is a combination of two words, namely demos and graphein which means people and writing. So demographics are any writings about the people or population [6]. Demographics are also a form of movement of population data, and the presence of demographic data will influence the decisions that will be taken by the existing government. Whereas residents are all people who have lived in the geographical area of the Republic of Indonesia for 6 months or more and or who have been domiciled for less than 6 months but aim to settle. On the government website, demographic figures are displayed, while for social conditions social demographics are used. The figures displayed are the authority of the existing government, as the website manager. Based on Law Number 32 of 2004 concerning Regional Government, and then became the basis for the implementation of regional autonomy that one of the mandatory functions carried out in the region is social affairs, including in the field of social welfare. The law is intended to bring government services closer to the objective needs of the community. In the context of the social welfare sector, so that those who face social welfare problems can be dealt with quickly and thoroughly. Through the regional autonomy policy, unnecessary burdens and tasks of the central government can be carried out by the regional government. This can be seen from the objectives of regional autonomy, namely: (1) improving services and improving community welfare, (2) developing democratic life, (3) social justice, (4) equity, (5) maintaining harmonious relations between the Center and regions and between regions in order integrity of the Republic, (6) push to empower communities and (7) foster initiative and creativity, increasing community participation, developing the role and functions of the Regional Representatives Council in the era of regional autonomy, local governments have been getting

* Corresponding authors
e-mail addresses: atmoko@usm.ac.id

reinforcement program from the center. In poverty reduction, through the deconcentration fund, the Ministry of Social Affairs has distributed programs and budgets for poverty reduction programs. Through strengthening these programs and budgets, it is hoped that the performance of regional governments will be more optimal. A poor population can be reduced from year to year. However, programs distributed from the center are supportive and do not take a greater role.

Based on the description above, the aim of this research is to look at providing aggregate information on social demographic data both from websites and through the role of the Social Service directly. And find out whether the existing data correlates with each other, as well as the relationship or influence of aggregate data on the perception of reading the data. It is hoped that the results of this research will provide material for increasing the role of social services in the field of research and data presentation.

2. Research Methods

This study uses the spatial analysis method of demographic figures displayed on government websites. Data and information are collected through the government website, study documentation by studying the results of previous research, reports and relevant literature. Data collected, in the form of aggregate data are then processed and analyzed. This study discusses and analyzes aggregate data for social-demographic variables that are measured spatially by using cross-correlation and Granger Causality causality tests. To represent spatial functions, social demographic aggregate data is collected from government websites. The method used is VAR / VECM bivariate (a model with two variables)[1]. One of the objectives is research to see whether there is a causal relationship between variables in social demographics displayed on government websites. Some tests used in this study include stationarity test, optimal lag selection, stability test, causality test, cointegration test, and VAR / VECM. Besides that, testing is also done to see the response of a variable as a result of shocks that occur in other variables (through the Impulse Response Function / IRF), and the contribution of a variable to changes experienced by other variables (through Forecast Error Variance Decomposition / FEVD). The data analyzed is first converted into natural logarithms. The use of natural logarithms shows that the slope coefficient X measures the constant proportions or relative changes in Y (the dependent variable) as a result of changes from X. In addition, the transformation of data into natural logarithms is done to obtain more valid results. Following is the flow of data analysis procedures used in this study.

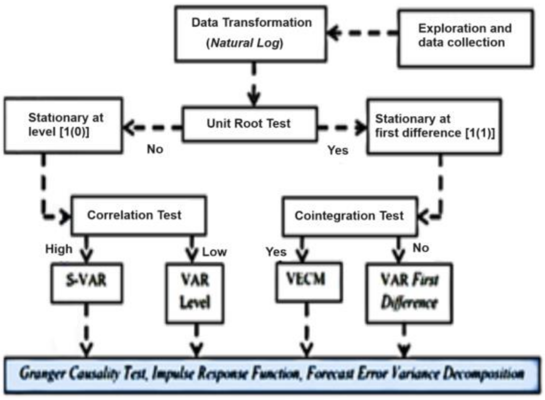

Data Transformation
(Natural Log)
Exploration and data collection
Stationary at level [1(0)]
Unit Root Test
Stationary at first difference [1(1)]
No
Yes
Correlation Test
Cointegration Test
High
Low
Yes
No
S-VAR
VAR Level
VECM
VAR First Difference
Granger Causality Test, Impulse Response Function, Forecast Error Variance Decomposition

Fig. 1. Flow of data analysis

The initial step in data analysis in this research is exploration and data collection. Researchers explore the aggregate data available on government websites. After exploring, the researcher collected the data in accordance with the objectives of this study. Next, the collected data is then converted into a natural logarithmic form (ln). After transforming the data, the first test phase is the stationarity test to see whether there is a unit root in the data used. Data is said to be stationary if there is no unit root, whereas if there is a unit root, it can be said that the data is non-stationary. When the unit root test results show that the data is stationary at the level, it can proceed to the next testing phase[5]. Conversely, when the data is non-stationary, it is necessary to do the next unit root test so that the data is stationary (it can be stationary at first difference or second difference). After the unit root test, the next step is to determine the optimal lag. Before determining the optimal lag, first, determine the maximum lag of a stable VAR model. After that, the optimal lag is determined based on the Akaike Information Criterion (AIC) criteria. Based on the unit root test stage at point c, it can be concluded that there is a possibility of data stationarity results, namely stationary at level [1 (0)] or stationary at first difference [1 (1)]. After conducting several stages of the above testing and decision making related to the analysis model used (S-VAR, VAR in level, VECM, or VAR indifference), then a causality test, Impulse Response Function (IRF), and Forecast Error Variance Decomposition (IRF), and Forecast Error Variance Decomposition (FEVD).

The approach to the Vector Autoregressive (VAR) model is built by minimizing theory (so it is often called a non-structural or non-theoretical model) so that economic phenomena can be captured properly [2][8]. And the VAR model can be divided into three forms, namely [2]:

1. Unrestricted VAR is a form of VAR that is related to cointegration problems and theoretical relationships where the data used is stationary data at the level. This VAR model is divided into two forms, VAR in level and VAR indifference. If the data is stationary at the level, VAR in level is used; whereas if the data is not stationary at the level and must be stationary before

the VAR model is formed, and does not have cointegration, then the estimated VAR indifference is used;

2. Restricted VAR, is a form of VAR that is often restricted and is often called the Vector Error Correction Model (VECM). It is called VECM because this model applies a gradual correction by means of a short-term adjustment to the deviation from the long-run equilibrium model. This model is used when there is cointegration on non-stationary data so it needs to be given a restriction on the long-term relationship of endogenous variables so that they converge into their cointegration, but still have short-term dynamics;

3. Structural VAR (theoretical VAR), is a form of VAR that restricts the variables used based on strong theoretical relationships and ordering schemes (ordering) relationship maps.

In his book, Juanda & Junaidi (2012) revealed that in a dynamic system, the relationship between variables cannot only be explained using a static single equation model but requires dynamic equations and mutual influence. This condition can create a VAR model (1), which can be written as the following Equation[2][4]:

$$Y_t = A_0 + AY_{t-1} + v_t \qquad (1)$$

And Juanda & Junaidi (2012) reveal that Equation 1 above is called the equation VAR 1 with two variables (bivariate) or in general, can be written as VAR (1). Juanda & Junaidi (2012) also revealed that if there are M variables observed with observations of T and order p, then the VAR model (p) can be written as follows[2][4]:

$$Y_t = A_0 + A_1Y_{t-1} + A_2Y_{t-2} + ... + A_pY_{t-p} + v_t \quad (2)$$

There are several stages of testing carried out before the VAR analysis, including stationarity test, determination of optimal lag, stability test, causality test, cointegration, and VAR formation / VECM, and analysis of Impulse Response Function (IRF) and Forecasting Error Variance Decomposition (FEVD).

Furthermore, stability testing is performed to ensure the stability of the model formed. Stability checking can be done using the polynomial function method or the roots of the characteristic polynomial. The stability of a VAR / VECM model can be seen from all the roots of the variables. If all roots have modulus that is smaller than one, then the model can be said to be stable. If all the roots of the polynomial function are inside the unit circle, the VAR model is considered stable and will produce a valid Impulse Response and Variance Decomposition analysis.

The concept of causality can be interpreted as a two-way relationship (not just one direction) between economic variables [7]. The relationship can be in the form of a causal relationship between variables. This test is used to see whether the performance of independent variable forecasting is influenced by independent variables. The concept of causality between two variables can be said as follows: "Variable X is said to cause Y (X granger cause Y) if the realization of X occurs earlier than Y, and if a regression of Y by including X gives significant results to predict Y". To see the causality of the VAR / VECM model, the Granger Causality Test can be used. The Granger Causality Test assumes that information relevant to predicting each variable is only found in time-series data[9][10].

Economically, the two variables are said to be cointegrated if there is a long-term relationship or balance between these variables. Cointegration is a long-term relationship between variables which, although not individually stationary, can be stationary when there is a linear combination between variables. It was also revealed that the cointegration test can only be done if the data used are integrated to the same degree. For example, if there are time-series data X and Y in which both data are stationary at the same level of differentiation [eg I (d)], then it can be said that the data are cointegrated. Variables that are proven to have cointegration must be included in the error correction model at the level, and if the variables are not cointegrated, they must be included in the VAR model at the level of differentiation. Cointegration can be tested in three ways namely Engle-Granger Cointegration Test, Cointegrating Regression Durbin Watson Test (CRDW), or Johansen Cointegrating Test.

3. Results and Discussion

Aggregate data used are data taken from the website of the Central Statistics Agency (BPS). Here an example is taken from the website https://semarangkota.bps.go.id/. And the aggregate data that will be taken are the poor population and the number of unemployed. The results are as follows:

Table 1. Number of poor people in Semarang City

Year	2011	2012	2013	2014	2015	2016	2017	2018	2019
Total	88453	83346	86734	84640	84270	83590	80860	73650	71969

From table 1 above shows the number of poor people in the period from 2011 to 2019, Table 2 shows the number of percentages of unemployment. The correlation between the number of poor people and the percentage of unemployed will be sought. From the aggregate data, it can be seen that the units used are different, the range of values is also different.

Table 2. Percentage of unemployment in Semarang

Year	2011	2012	2013	2014	2015	2017	2018	2019
Total	7.65	6.01	6.02	7.76	5.77	6.61	5.29	4.54

Interestingly, the data shown in table 2 are not sequential, in 2016 there were none, so if correlations had to be found the values had to be compared, with Matlab the following results were obtained:

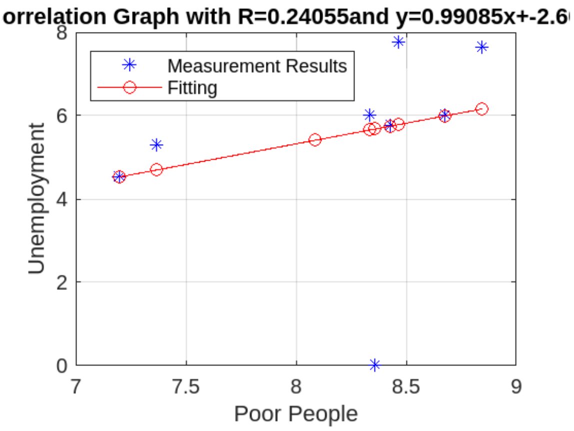


Fig. 2. Correlation with non-sequential data

From Figure 2 shows no correlation, with R values far from 1, if we fix for the missing data (preprocessing), the results will be as follows:

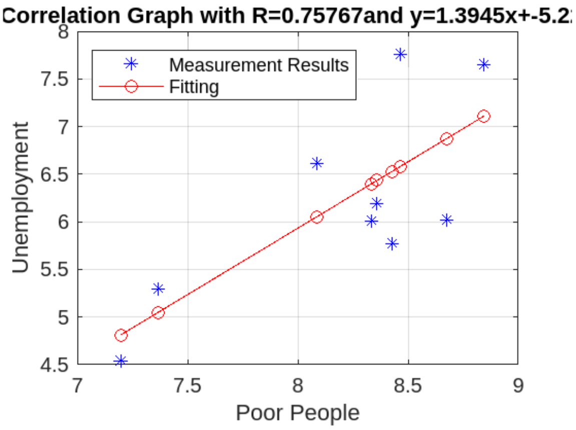


Fig. 3. Correlation with the improved data

Figure 3 looks at R = 0.76 close to 1 so that it can be said if there is a correlation between the poor and unemployed, although from the data displayed on government websites are often found incomplete data and data that are not the same in representing it (poor people in number and unemployment in percentage).

Figure 4 displays a graph between poor people and unemployment. From this picture, it can be seen that these two time series are indeed correlated, but do they influence each other? We can test this with the Granger causality test.

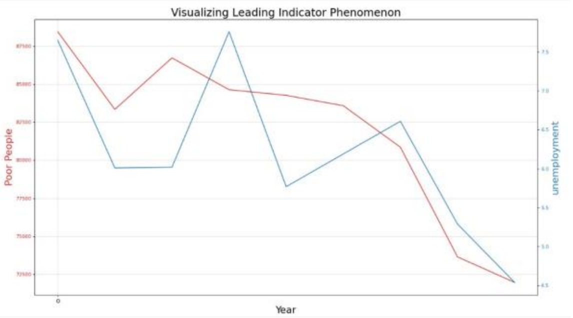
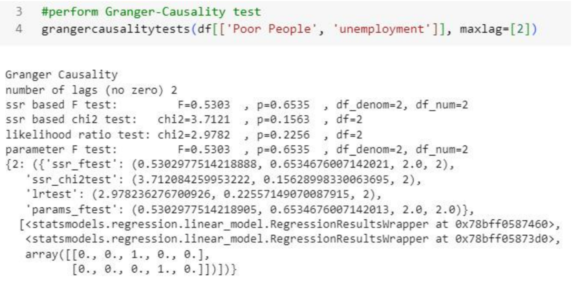


Fig. 4. Visualizing poor people and unemployment

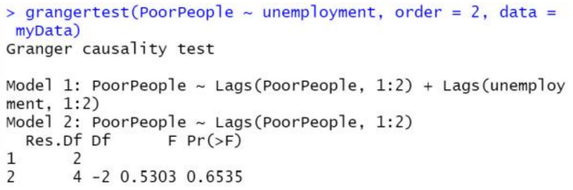
The Granger causality test is employed to assess the potential of one time series to serve as a predictor for another. The null hypothesis (H0) posits that there is no causal relationship between time series X and time series Y, specifically in the context of Granger causality, where time series X does not have the ability to Granger-cause itself. The alternative hypothesis (H1) posits that there is a causal relationship between time series X and time series Y, where X Granger-causes Y. The concept of "Granger-causes" refers to the valuable ability to anticipate the value of a time series Y in a future time period based on knowledge of the value of a time series X at a specific lag. With Python and R, the following results were obtained:



```
#perform Granger-Causality test
grangercausalitytests(df[['Poor People', 'unemployment']], maxlag=[2])
```

```
Granger Causality
number of lags (no zero) 2
ssr based F test:         F=0.5303  , p=0.6535  , df_denom=2, df_num=2
ssr based chi2 test:   chi2=3.7121  , p=0.1563  , df=2
likelihood ratio test: chi2=2.9782  , p=0.2256  , df=2
parameter F test:         F=0.5303  , p=0.6535  , df_denom=2, df_num=2
{2: ({'ssr_ftest': (0.5302977514218888, 0.6534676007142021, 2.0, 2),
   'ssr_chi2test': (3.712084259953222, 0.15628998330063695, 2),
   'lrtest': (2.978236276700926, 0.22557149070087915, 2),
   'params_ftest': (0.5302977514218905, 0.6534676007142013, 2.0, 2.0)},
  [<statsmodels.regression.linear_model.RegressionResultsWrapper at 0x78bff0587460>,
   <statsmodels.regression.linear_model.RegressionResultsWrapper at 0x78bff05873d0>,
   array([[0., 0., 1., 0., 0.],
          [0., 0., 0., 1., 0.]])])}
```

Fig. 5. Granger Causality Test in Python



```
> grangertest(PoorPeople ~ unemployment, order = 2, data =
  myData)
Granger causality test

Model 1: PoorPeople ~ Lags(PoorPeople, 1:2) + Lags(unemploy
ment, 1:2)
Model 2: PoorPeople ~ Lags(PoorPeople, 1:2)
  Res.Df Df      F Pr(>F)
1      2
2      4 -2 0.5303 0.6535
```

Fig. 6. Granger Causality Test in R

From Figures 5 and 6, it is shown that the F test statistic is denoted by the letter F, which equals 0.5303, and the p-value that corresponds to the F test statistic is Pr(>F) = 0.6535. This result agrees with the null hypothesis (H0) because the p value is greater than 0.05 and concludes that the unemployment value is not useful (has no effect) for predicting the value of poor people in the future. This is possibly caused by the unemployment value being a percentage, not a number.

4. Conclusion

From the discussion obtained to test the causality of aggregate data on government websites, it is possible that if there are similar perceptions, such as poor people and unemployment, correlation values can be obtained. However, it is also important to pay attention to the fact that the data is presented sequentially, that nothing is lost over a certain period of time, and that it is not in the form of aggregates such as percentages.

5. References

[1] Alhussayen, Hanan. "The Relationship Between Trading Volume and Market Returns: A VAR/Granger Causality Testing Approach in the Context of Saudi Arabia." Organizations and Markets in Emerging Economies 13.1 (2022): 260-276.

[2] B. Juanda and Junaidi, Ekonometrika deret waktu teori dan aplikasi. 2012

[3] M. D. Anggreani, E. P. Purnomo, and A. N. Kasiwi, “Ruang Publik Virtual Sebagai Pintu Komunikasi (Studi Kasus : Perbandingan Media Sosial Pemerintah Kota Yogyakarta dan Surabaya),” vol. 6, pp. 203–220, 2020

[4] Massa, Ricardo, and Juan Rosellón. "Linear and nonlinear Granger causality between electricity production and economic performance in Mexico." Energy Policy 142 (2020): 111476.

[5] Muritala, Taiwo A., et al. "Fraud and bank performance in Nigeria–Var granger causality analysis." Financial Internet Quarterly 16.1 (2020): 20-26.

[6] N. Fitriani, T. Militina, and A. S. Effendi, “Pengaruh Faktor Demografi Dan Investasi Swasta Terhadap Pertumbuhan Ekonomi Kota Samarinda,” J. Ekon. Pembang., vol. 10, no. 1, p. 47, 2012.

[7] S. E. Kapahang, E. Mingkid, and E. R. Kalesaran, “Keterbukaan Informasi Publik Pada Dinas Kominfo Pemerintah Kabupaten Minahasa Tenggara,” -.

[8] Shojaie, Ali, and Emily B. Fox. "Granger causality: A review and recent advances." Annual Review of Statistics and Its Application 9 (2022): 289-319.

[9] Su, Yong, et al. "Does tourism affect economic growth of China? A panel granger causality approach." Sustainability 13.3 (2021): 1349.

[10] Sunde, Tafirenyika. "Energy consumption and economic growth modelling in SADC countries: an application of the VAR Granger causality analysis." International Journal of Energy Technology and Policy 16.1 (2020): 41-56.