2020 Asia–Pacific Statistics Week
A decade of action for the 2030 Agenda: Statistics that leaves no one and nowhere behind
15-19 JUNE 2020 | Bangkok, Thailand

Development of Integrated Analysis with the Interpolation to Small Area Model in Inflation data (CPI)

PARAMARTHA, Dede Yoga[1]; MAJIDAH, Anisa Muna [1]

[1] Badan Pusat Statistik

Abstract:
The distruptive era raises hopes and challenges for the national statistics office, and BPS is no exception. In the era of distruptive information, many analyzes that can be developed do not close to complex analyzes such as stochastic spatial Interpolation. Spatial interpolation which usually can present SDGs indicators to the smallest stage can be facilitated by processes with the existence of Big Data as auxiliary variables. The benefit of using big data to spatial interpolation is using auxiliary variable with no standard errors. Big data will indeed have a bias in its calculations, but there is big data that is able to provide good ordinality to the smallest level of disaggregation intended. The examples of big data that can be utilized are night time light (NTL) intensity data and also the number of e-commerce accounts estimated at the municipal district level. NTL data which provides the concept of socioeconomic activity from an area at night combined with data on the number of accounts in e-commerce will be able to be used to proxy inflation at district level. In this study, inflation estimation from the district level is also used in meeting indicators that meet the SDGs objectives, effectiveness of monetary policy. In this study, the resulting interpolation rate of inflation at the district level on the island of Java. By using big data as an auxiliary variable, researchers conducted a correlation test of results with other indicators that have economic closeness to inflation to provide supporting evidence of the potential of big data as an auxiliary variable so that in the future, the use of big data can facilitate the BPS business process.
Keywords: Big Data, Spatial-Interpolation, Disaggregation, Monetary Policy.

1. Introduction:

One of the macroeconomic indicators used to see the stability of a country's economy is inflation. Inflation shows a tendency to increase prices of goods and services in general, which continues over time. The inflation rate is measured through the Consumer Price Index (CPI) indicator. The CPI is an index that calculates the average price change of a package of goods and services consumed by households in a certain period of time. Changes in the CPI over time reflect the rate of increase (inflation) or the rate of decline (deflation) of goods and services.

Until now, inflation in Indonesia was not available for all regional units. Starting in January 2020, inflation covered 90 cities, an increase from before which covered 82 cities. This is based on the 2018 Cost of Living Survey conducted in 90 cities, consisting of 34 provincial capitals and 56 districts / cities. Cost of Living Survey is used as a weight in the preparation of CPI.

The limited area covered by the inflation calculation is a constraint. This causes conditions in other cities to be invisible. Therefore, describing inflation in all regions is an important thing to do, to obtain a more detailed depiction of the region. In addition, this can also be an explanation if an anomaly occurs. Related to SDGs, it can be used in the preparation of misery index, which is an index used to measure the economic conditions of a country, which is a combination of the inflation rate (CPI) and the unemployment rate.

Another spatial benefit is that it can be used as a mapping of vulnerable areas if a crisis will occur. With the availability of more detailed depictions in all regions, it is possible to find out more specific regional characteristics. In addition, it can illustrate the potential of the region that has not been previously described.

Utilization of big data is one of the breakthroughs in meeting the growing needs of data. Support from the broader big data can be optimized in the context of meeting the needs related to the availability of inflation in all regions in more detail. Big data that can be used include night time light (NTL) and data from one of the top e-commerce sites in Indonesia called shopee. From the collaboration, it is expected to produce more specific indicators, which later can be used to compile indexes, and develop more advanced methodologies.

2. Methodology:

Inflation

Inflation is a rising price of goods and services in general that continues over time. The inflation rate is calculated from the Consumer Price Index (CPI). The CPI calculates the average price change of a package of goods and services consumed by households in a certain period of time. Changes in the CPI from time to time indicate the level of inflation/deflation of goods and services. Some theories that are often used in explaining inflation include the quantity theory, the Keynesian theory, the "cost-push" theory, and the structural theory. First is the quantity theory which explains that the price level is determined by the amount of money. The price level will be directly proportional to the amount of money and inversely proportional to the physical volume of production. The second is the Keynesian theory. This theory assumes that consumers tend to spend a fixed proportion of every increase they receive in their income. Inflation arises entirely from efforts to buy more goods and services than can be provided, or more than can be produced at the level of "full employment" activity. The third is the "cost-push" theory. This theory assumes that prices of goods are basically determined by their costs, whereas money supply is responsive to demand. Finally, the structural theory, which emphasizes structural mismatches in the economy.

E-commerce

E-commerce is the buying and selling of goods and services on the internet. E-commerce is rapidly changing the way companies interact with each other as well as with consumers and government. As stated in UNCTAD (2015) as a result of changes in ICT, e-commerce is currently developing rapidly in developing countries. Developments in the internet have created new opportunities for e-commerce and created a new set of global and national trade. This allows many buyers and sellers to be on a common platform and results in increased efficiency.

Economic big data

The concept of big data first appeared in the late 90s by Cox and Ellsworth (1997) and was defined in the early 2000s by Laney (2001) as 3Vs namely Volume (data size), Velocity (data transfer rate), and Variety (various types of data). This model was developed and expanded to 4V, with the addition of the Value (the process of extracting valuable information from data). At present, big data is defined in 5Vs by Bello-Orgaz et al. (2016) with the added dimension of Veracity (related to proper data management and privacy issues). The big data paradigm offers many advantages and benefits as stated by Jin et al. (2015). The contribution of big data is very potential for national and industrial development, because it encourages change and improvement of research methods, as well as forecasting.

Big Data That Used *(Data source)*

Night time light (NTL) data was obtained from the U.S. Air Force's Defense Meteorological Satellite Program/Operational Linescan System (DMSP/OLS). DMSP satellites produce global night time and day time which covers the earth every 24 hours with the main objective to monitor the distribution of clouds and assess navigation conditions. The U.S. Department of Commerce's NOAA National Geophysical Data Center (NGDC) takes data from the DMSP satellite and performs extensive processing and includes an algorithm to produce a night time light emission database.

Data on the number of shopee accounts is obtained through crawling at the address shopee.co.id. The tool used is python. The package used in python is the scrapy package. The data taken is data at one point of time, namely in November 2019. The steps of the scrapping carried out begins with taking a list of user data from the shopee on the shopee sitemap. After obtaining all shopee users, a filter is then performed to determine the user category whether the store or individual. If the user is categorized as a store, then the next will be scrapping information in the form of a store location. Data that has been extracted is then aggregated into a tabulation of the number of shops in regencies/cities throughout Java Island.

Henceforth in modeling, the data with the night time light specification will be named NTL, and for auxiliary numbers the shopee account will be stated with N.Shopee.

Kriging Regression

Spatial interpolation can be defined as the process of estimating the value of a function that has real value in $z(s)$ based on the set of values $z(z(s_1), z(s_2), \dots, z(s_n))$ at the different points $\{s_1, s_2, \dots\ s_n\}$. The deterministic models such as Inverse Distance Weight, RBF, and spline are widely applied to spatial interpolation. However, a deterministic approach may not be appropriate because in geography, environmental science, and ecology we usually lack sufficient information about the variation of properties in space (Meng, et al. 2013). Thus, there needs to be a stochastic approach in accommodating spatial random effects, where one of them is by using kriging regression.

In the spatial interpolation process, suppose the sample value for a given location is represented as $z(s_1), z(s_2), z(s_3), \dots, z(s_n)$, where $s_n = (x_{latitude}, y_{longitude})$ is the location point with $x_{latitude}$ and $y_{longitude}$ coordinates, with location $i = 1, 2, 3, \dots, n$. Values at new and non-amputated locations (s_0) can be predicted using RK by adding spatial trends and random components (i.e. residuals) (Odeh et al. 1994) using the following equation:

$$\hat{z}(s_0) = \hat{m}(s_0) + \hat{e}(s_0)$$

where $\hat{e}$ residue is interpolated using ordinary kriging and the trend is suitable using linear regression as follows:

$$\hat{z}(s_0) = \sum_{k=0}^{p} \hat{\beta}_k \cdot q_k(s_0) + \sum_{i}^{n} w_i(s_0) \cdot e(s_i)$$

where $\hat{\beta}_k$ is the estimated regression coefficient kth, the external auxiliary variable k-th at location s_0 is q_k (temporarily, $q_k(s_0) = 1$), the number of auxiliary variables is indicated as p, the weight is $w_i(s_0)$, and the regression residue is $e(s_i)$. The RK can represent in matrix notation as follows:

$$\boldsymbol{z} = \boldsymbol{q}^T \cdot \boldsymbol{\beta} + \boldsymbol{\varepsilon}$$
$$\hat{z}(\boldsymbol{s_0}) = \boldsymbol{q_0^T} \cdot \widehat{\boldsymbol{\beta}} + \boldsymbol{\lambda_0^T} \cdot \boldsymbol{e}$$

where $\boldsymbol{\varepsilon}$ is the regression residue, q_0 is the vector of the auxiliary variable p at $\boldsymbol{s_0}$, $\widehat{\boldsymbol{\beta}}$ is a vector of the estimated coefficient of model $p + 1$, $\boldsymbol{\lambda_0}$ is a vector of n kriging weights, and $\boldsymbol{e}$ is a vector of n residues. Accounting for residual spatial correlations, we can solve the regression coefficients using the following general least squares estimate (Cressie 1993):

$$\hat{\beta} = (q^T \cdot C^{-1} \cdot q)^{-1} \cdot q^T \cdot C^{-1} \cdot z$$

where q is a matrix of auxiliary variables at all observed locations, z is a vector of observations of sample responses, and C is a covariance matrix n × n of residuals.

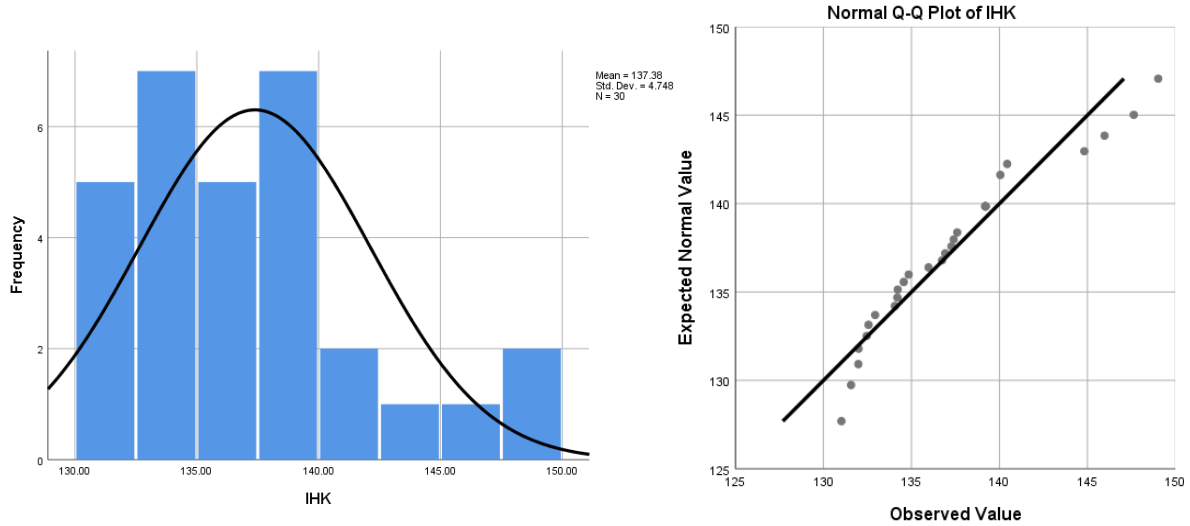
Examples to illustrate the kriging regression process are to do simple or multiple linear regression, select an effective semi-variogram model, and then use it to model residuals, calculate regression coefficients, process weighting matrices, and finally use kriging regression models, and get predictions at points that are not interfere (Meng, et al. 2009). In the RK process, the regression residue (i.e. the uncertainty model) is entered into the interpolation system, which is then applied directly to the spatial prediction of the primary variable.

Data Specification

The data used are data with a period of 2019. The unit of analysis and observation are all municipal districts in Java. The approach to inflation is proxied by the CPI. CPI is the foundation of inflation calculation, and if it can be interpolated, then inflation should be calculated by using interpolation at two time points. In addition, the weight used in kriging regression is distance inverse weight as a representative of the spread effect of commodity prices.

3. Result:

Before proceeding further, the analysis begins with in-sample data exploration of the variables to be interpolated. After this has been done, it is expected that sufficient results will go to the next process, namely spatial interpolation. The results of the initial exploration can be described in Figure 1.



Kolmogorov-Smirnov Test (n=30)	Test Statistic	0.151
	Asymp. Sig. (2-tailed)	0.080

Figure 1. Distribution of CPI data in the In-sample area

It appears that the CPI data has a histogram that tends to have a fairly normal distribution. However, there may be several outliers at the top value. Turning to analysis with Q-Q plots, it appears that most CPI values are already at the origin of the normal distribution. Finally tested with statistical analysis. Based on the analysis using the Kolmogorov-Smirnov test, it was found that at the 5% significance level, the CPI data has a normal distribution.

With this signal, then interpolation can be done at a later stage with kriging regression. Even with GLM estimation, kriging regression requires normal assumptions for the validity of the interpolation results. With the fulfillment of the normal distribution, the interpolation results can be said to be quite statistically valid.

GLM Model for Kriging Regression

After fulfilling the normality assumption, the next step is to look for parameters and residual vectors of the two models with auxiliary variables to be compared and their capabilities tested. The results of the model specifications can be seen in the following table.

Model by	Intercept	β estimate	Residual Deviance	df Resid
NTL Aux.	135.22**	0.0378 *	531.4	28
N.Shopee Aux.	136.20*	2.32.E-05*	588.4	28

**) 1% significant ; *) 5% significant

Table 1. Model Specification with GLM-estimated

The results show that both models can validity be used with the significance of the parameters used. From both models, it appears that models with auxiliary NTL will have lower variations compared to models with auxiliary N.shopee. Based on the intercept value it can also be seen that the NTL model has a lower origin.

Model Validation

Model by	MAPE (%)	MSE	AIC	MAE
NTL Aux.	0.2334	0.5203	177.40	0.3293
N.Shopee Aux.	0.2855	0.5559	180.40	0.4020

Table 2. Model Validation Statistics for In-sampled Data

Switching to how like big data is used as an auxiliary variable. It can be seen from the MAPE values of the two models that are very low with a value of less than 5%. This shows that the predictions made in

the areas sampled for CPI calculations can be estimated well by the big data used. This is also due to the theoretical foundation of selecting big data that is relevant to be used as an auxiliary variable. Furthermore, when compared, the NTL model appears to have better validity compared to the N.Shopee model. This can be seen from all the validity indicators used, which all show the superiority of the NTL model. However, to determine the model used there needs to be validity in terms of mapping and distribution in order to prove it is in accordance with the facts that occur.

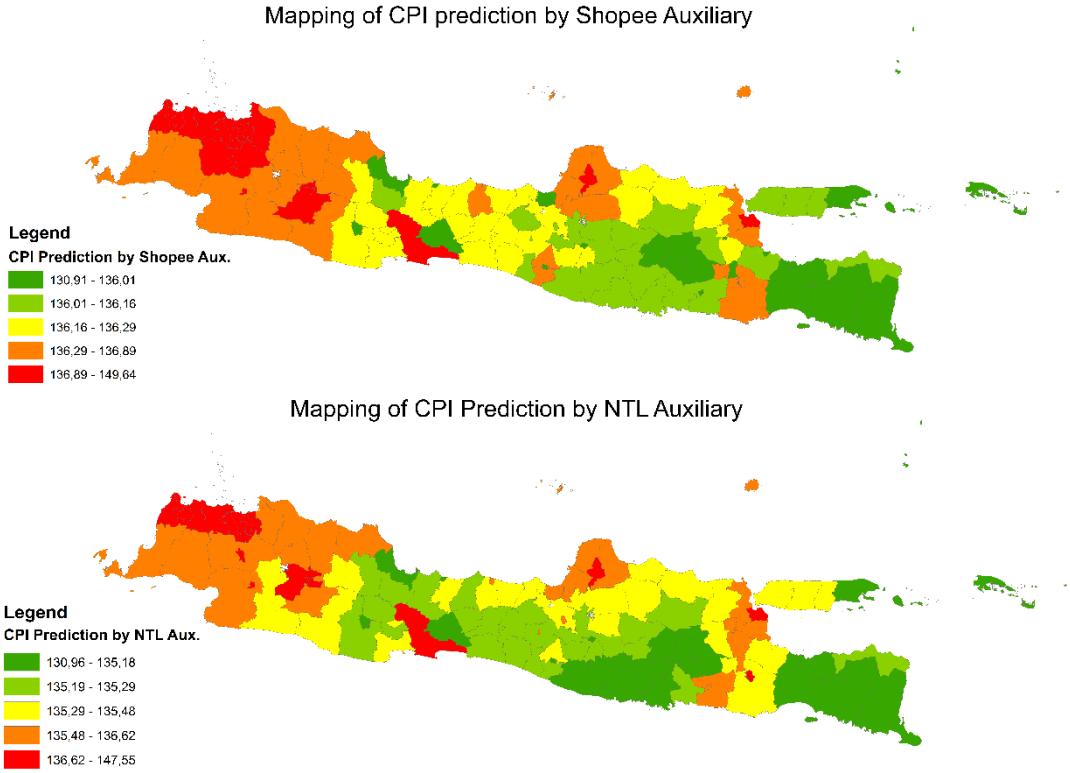


Figure 2. Estimated CPI Data Distribution by GIS mapping.

Figure 2 is a GIS mapping with a cut-off for categorizing the quantile method. Based on the description of the two models, the spatial distribution has a similar pattern, where the western region has a high CPI. What is even more differentiating is that the N.Shopee model further illustrates how provinces are divided by their consumer price indexes. Meanwhile, the NTL model looks more at how interpolation is the spread effect to the surrounding area.

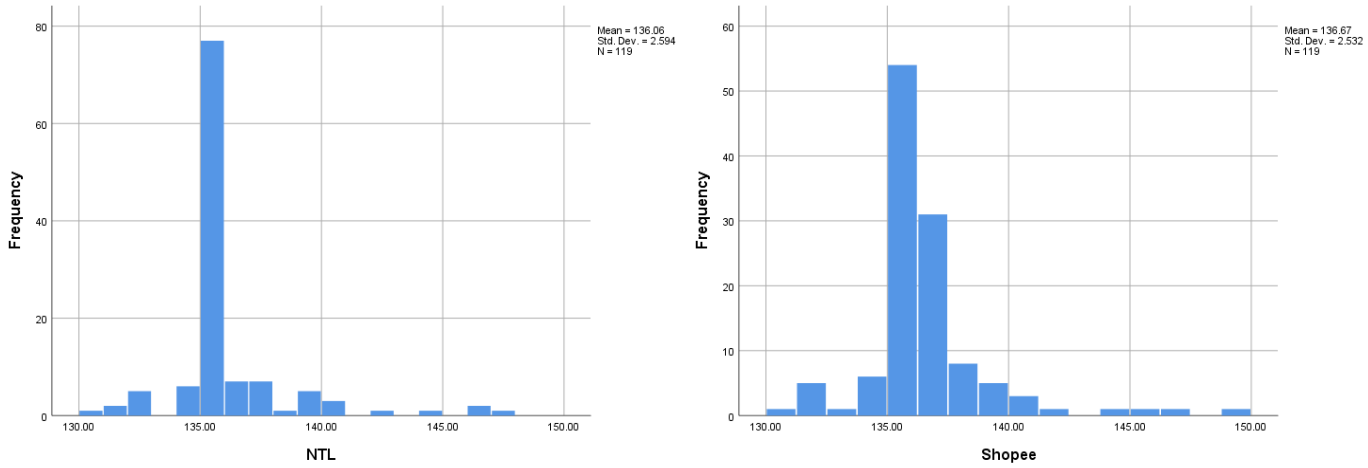


Figure 3. Distribution of Estimated CPI Data with Histogram.

The histogram results in Figure 3 also show relatively similar patterns for the two models. However, in terms of distribution, although both peaked at the midpoint, the empty shopee model is not too extreme. This is seen from the peak which has a frequency of about 50% of all data.

Findings

From the results of the analysis, it was found that the results were quite representative of the cities and regencies. This can be seen from the distribution pattern of CPI in figure 2 which shows the high CPI is in urban areas on the island of Java. In addition, this is indicated in the western regions. In the western regions, it is seen that the CPI value is very high, this is in view of the existence of Jakarta which is the economic center as well as the Indonesian government. In addition, regions with high economic traffic in the western region are also supported by the spread of effects by Jakarta and Tangerang. This is sufficient proof of validity to prove the high CPI in the western region compared to other regions on the island of Java.

Apart from validity based on available facts, on the other hand, big data can be used as an alternative to interpolating with the resulting pattern. Coupled with high data rates, big data can always be an alternative in becoming an integrated statistic with quite good speed and accuracy. In addition, more up-to-date data can be used as an acceleration of several statistical indicators in compiling faster and specific indicators, such as misery indexes for depicting SDGs.

4. Discussion, Conclusion and Recommendations:

This study found that big data as a new alternative. Later big data will be used more often by having sufficient accuracy to describe facts in the field and theoretically, even though big data still has problems in data pre-processing.

The use of big data in the future can also support the monitoring of SDGs indicators. It can be seen from this study that is able to describe the condition of the consumer price index in the district/city.

References:

1. Bello-Orgaz, G., Jung, J.J., Camacho, D., 2016. Social big data: recent achievements and new challenges. Inf. Fusion 28, 45–59. http://dx.doi.org/10.1016/j.inffus.2015.08. 005
2. Cox, M., Ellsworth, D., 1997. Managing Big Data for scientific visualization. ACM Siggraph, MRJ/NASA Ames Res. Cent. 5, 1–17.
3. Cressie, N. 1993. Statistics for Spatial Data, Revised Ed. New York: Wiley.
4. Jin, X., Wah, B.W., Cheng, X., Wang, Y., 2015. Significance and challenges of big data research. Big Data Res. 2 (2), 59–64. http://dx.doi.org/10.1016/j.bdr.2015.01.006.
5. Laney, D., 2001. 3D Data Management: Controlling Data Volume, Velocity, and Variety. Application Delivery Strategies. pp. 949. http://blogs.gartner.com/doug-laney/files/ 2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-andVariety.pdf
6. Qingmin Meng , Zhijun Liu & Bruce E. Borders (2013): Assessment of regression kriging for spatial interpolation – comparisons of seven GIS interpolation methods, Cartography and Geographic Information Science, 40:1, 28-39.
7. UNCTAD (2015). Information economy report 2005: Unlocking the Potentials of e-commerce for developing countries. United Nations Publication.