WWW.IJRPR.COM

International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Geographically Weighted Principal Component Analysis Gaussian to Reducing Factors Affecting Happiness

Anggun Yuliarum Qur'ani, Ratna Sari Widiastuti [a]

[a] *Faculty of Mathematics and Natural Science, Udayana University, Bali 80361, Indonesia*

A B S T R A C T

Principal Component Analysis (PCA) is often used by researcher to reduce the dimensionality of a dataset. PCA that involves spatial weighting with a point approach is called Geographically Weighted Principal Component Analysis (GWPCA). Geographically Weighted Principal Component Analysis (GWPCA) Gaussian is GWPCA that uses Gaussian Kernel Adaptive Weighting. Based on the results of the Gaussian GWPCA, 26 components are formed, but 3 components are sufficient to represent the diversity of the data, which is equal to 83.5%. The main factor that influences the highest happiness of the 26 Indonesian citizens is marriage with a family of 4 people, and this is felt more by the wife.

Keywords: GWPCA Gaussian, Spatial Weighting, Happiness Index, Indonesian Citizens.

Introduction

Geographically Weighted Principal Component Analysis (GWPCA) is a development of Principal Component Analysis (PCA) that involves location weighting with a point approach. Comber, et al (2016) said that GWPCA is a local adaptation of PCA that locally transforms image data, and as such, can describe spatial changes in the structure of multi-band images, thus directly reflecting that many vertical processes are spatially heterogeneous. Gaussian Geographically Weighted Principal Component Analysis (GWPCA) is a GWPCA that uses a Gaussian spatial weighting matrix. . The variable X_i at location i is assumed to have a Multivariate Normal distribution with vector mean $\mu(u,v)$ and Matrix variance-covariance $\Sigma(u,v)$ with (u,v) is the spatial location i. So, we have $X_i \sim N(\mu(u,v);\ \Sigma(u,v))$.

Ayundanisa, et al. (2021) used GWPCA to cluster data on regional revenue sources for districts and cities in West Java in 2021. They get 3 main components with an accuracy rate of data variation of 80%. Mas'ad, et al (2016) have conducted an analysis using GWPCA to cluster the factors that affect the percentage of poor people in Central Java. They get the accuracy of the clustering model into 3 main components of 87.2499%. In addition, Comber, et al (2016) used a reference data set for land cover west of Jakarta, Indonesia, with a classification procedure assessed through an 80/20 split of training and validation data, repeated 100 times. For each classification algorithm, the inclusion of GWPCA weighting data was found to significantly improve classification accuracy.

Al (2019) used panel regression analysis method to look at the variables that affect the happiness index in Indonesia in the period of 2014 and 2017. He got the results of analyzing 1 general model for all provinces with an accuracy of 99.07%. Dewi (2020) used the multipolynomial logistic regression method to determine the factors that affect the level of happiness with the modeling results explaining the accuracy of the proposed model is only 4%.

From this background, researcher want to reduce the variables that affect the happiness index of Indonesian citizens in 2021 using GWPCA Gaussian so that the proposed model is correct and the accuracy of the model can be known for each province in Indonesia.

Literature Review

LeSage et al (2009) say that spatial dependence (autocorrelation) reflects the situation that is observed at one location depends on the value of neighboring observations at nearby locations. One method used to detect the presence of spatial autocorrelation (Nuroini, 2019) is the Moran's I test with the hypothesis:

$H_0: I = 0$ (there is no spatial autocorrelation)

$H_1: I \neq 0$ (there is spatial autocorrelation)

Statistics test of Moran's I:

$$I = \frac{n\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}(x_i-\bar{x})(x_j-\bar{x})}{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}\sum_{i=1}^{n}(x_i-\bar{x})^2}, i \neq j \qquad (2.1)$$

Where I: Moran index; n: number of observations; x_i: observed value of the variable at location$-i$, $i = 1,2, \dots, n$; x_j:observed value of the variable at location$-j$, $j = 1,2, \dots, n$; w_{ij}: elements in row$-i$ and column$-j$ of the spatial weight matrix. If H_0 is false, then the test of statistic $(I) = \frac{I-E(I)}{\sqrt{\sigma^2_{(I)}}}$ is rejected which is compared to the Normal distribution for $E(I)$ and $\sigma^2_{(I)}$ (Pebesma, et al., 2023), that is $Z(I) > Z_{\alpha/2}$. So, there is spatial autocorrelation.

Spatial weighting is needed to represent the location of observation data with one another. In this study, Adaptive Kernel Gaussian weighting is used which can be formulated (Qur'ani, 2014) as follows:

$$w_j(u_i,v_i)=(-0,5)\left(\frac{d_{ij}}{h}\right)^2 \quad (2.2)$$

where $d_{ij} = \sqrt{(u_i - u_j)^2 + (v_i - v_j)^2}$ in equation (2.2) is the Euclidean distance between location (u_i, v_i) to location (u_j, v_j), and h is a non-negative parameter that is known and usually referred to as the smoothing parameter (bandwidth). The selection of the optimum bandwidth is very important because it can affect the results of model estimation (Safitri, 2021). To get the optimum bandwidth, it can be done by calculating Cross Validation (CV). If the CV value is getting smaller, the optimum bandwidth is obtained (Fotheringham et al, 2002) using the following formula:

$$CV = \sum_{i=1}^{n} [x_i - \hat{x}_{\neq i}(b)]^2 \quad (2.3)$$

where i is the i −th location, b is the bandwidth, is the value of the observation without the i −th observation. Other kernel shapes can also be used in the GW model. The GW principal component for the location (u_j, v_j) (Harris, 2011) can be formulated as

$$LVL^T|(u_i, v_i) = \Sigma(u_i, v_i) \quad (2.4)$$

where $\Sigma(u_i, v_i)$ is the variance-covariance matrix of GW for location (u_i, v_i).

Methodology

The methods in this study explain the data sources, and data analysis methods. The data analyzed using GWPCA Gaussian is secondary data taken from the BPS Happiness Index in 2021 in 34 provinces. The data consists of components that form elements of the measurement of the happiness index in urban areas from 34 provinces in Indonesia in 2021, consisting of 28 factors.

The steps taken in this analysis are as follows.

- Testing the assumption of multivariate normality, and the assumption of spatial autocorrelation.

- Calculating the optimum bandwidth using Equation (2.3)

- Determining the Gaussian spatial weight matrix in Equation (2.2).

- Perform GWPCA. In this study, the analysis used R software, package GWModel in Harris et al (2015), function "bw.gwpca".

- Interpretation of results and discussion.

Results and Discussion

The multivariate normality assumption test using the Shapiro-Wilk test for Multivariate Normal on 28 factor data taken from (BPS, 2021) gives p-value is 0.1181 which is shown in Table 1.

Table 1 Multivariate Normal Assumption Test Results

Statistics Test	MVN value	p-value
MVW	0.9641	0.1181

So the data is multivariate normally distributed.

For the assumption of spatial autocorrelation, it can be shown in full in Table 2 below.

Table 2 The p value of the Moran's I test

Factor	P value
Single	**0.0949538**
Marriage	4.54E-07
Living Divorce	5.29E-05
Death Divorce	9.13E-08
Primary School to 24 years old	0.0026328
25 to 40 years old	9.28E-07
41 to 64 years old	5.11E-07

Factor	P value
> 65 years old	2.74E-05
Husband	2.85E-06
Wife	4.17E-07
1 person at home	0.0001375
2 person at home	2.87E-05
3 person at home	2.02E-06
4 person at home	1.64E-06
5 person or more at home	1.10E-06
Never attended school	1.84E-05
Not graduated from Primary School	3.84E-07
Primary School	1.19E-08
Elementary School	4.82E-06
High School	4.45E-07
Diploma I/II/III	9.37E-08
Diploma IV or Bachelor	0.0033942
Master or Doctoral	**0.8554286**
Salary < 1.8 million	4.84E-08
Salary 1.8 to 3 million	1.44E-08
Salary 3 to 4.8 million	6.18E-08
Salary 4.8 to 7.2 million	9.94E-07
Salary >7.2 million	5.83E-08

Source: Assumption test using R software

Where $\alpha = 0{,}05$, there are 2 factors whose values are greater than α, which means there is no spatial autocorrelation. Thus, these 2 factors are not involved in GWPCA Gaussian. The optimum bandwidth value by looking at the smallest Cross Validation (CV) value is 221067.8.

Descriptively, the happiness index of Indonesian citizens in 2021 can be mapped as Figure 1.

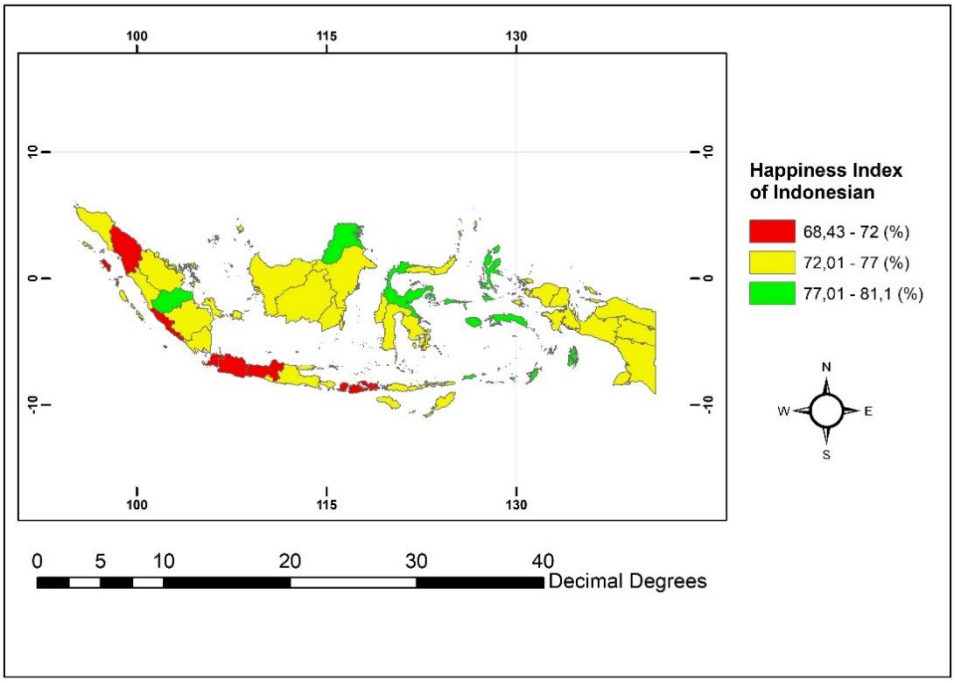

Figure 1: Happiness Index Map of Indonesian Citizens in 2021

Furthermore, from the results of the GWPCA Gaussian analysis, the 3 PCA components represent 83.25% of the data diversity. However, component 1 already represents 74.51% of the data diversity shown in Table 3.

Table 3. Components formed from GWPCA Gaussian

Statistics	Component 1	Component 2	Component 3
Standard deviation	4.568	1.141	1.071
Proportion of Variance	0.7451	0.0465	0.0409
Cumulative Proportion	0.7451	0.792	0.8325

Source: GWPCA Gaussian analysis using R software

Based on the loading results for Component 1 to Component 3 in Table 4. Indonesians in Bali Province have the highest main factor affecting happiness is education. In fact, those who feel the most happy are Indonesians who have never attended formal education at school. This is not only the case in Bali Province, but also in all provinces in Indonesia.

The second factor and several factors that have a similar effect on the happiness of Indonesians is that Indonesians over 65 years old who are living alone in their homes, whether they are living divorce or death divorce are happier. For component 1, all the coefficients of all the factors have a positive effect, increasing the effect on the happiness of Indonesians.

Table 4. Loading coefficients formed by PC1. PC2. and PC3 for Bali Province

Factor	Component 1	Component 2	Component 3
Marriage	0.184	0.011	0.057
Living Divorce	0.201	-0.085	-0.205
Death Divorce	0.216	-0.038	-0.176
Primary School to 24 years old	0.195	-0.012	-0.181
25 to 40 years old	0.175	0.019	0.127
41 to 64 years old	0.186	-0.036	-0.033
> 65 years old	0.220	0.081	-0.132
Husband	0.188	-0.037	0.015
Wife	0.191	0.069	0.034
1 person at home	0.215	0.028	-0.304
2 person at home	0.195	0.005	-0.053
3 person at home	0.198	-0.015	0.058
4 person at home	0.195	0.040	0.029
5 person or more at home	0.171	-0.026	0.097
Never attended school	0.300	0.291	-0.101
Not graduated from Primary School	0.170	-0.027	-0.146
Primary School	0.200	0.029	0.012
Elementary School	0.178	0.002	0.054
High School	0.168	0.015	0.087
Diploma I/II/III	0.156	0.146	0.701
Diploma IV. Bachelor	0.175	-0.002	0.231
Salary < 1.8 million	0.197	0.014	-0.198
Salary 1.8 to 3 million	0.196	-0.012	0.012
Salary 3 to 4.8 million	0.177	-0.010	0.081
Salary 4.8 to 7.2 million	0.140	-0.006	0.184
Salary >7.2 million	0.130	-0.120	0.264

Source: GWPCA Gaussian analysis using R software

Looking at Component 1 of marital status, Indonesians feel happier when their marital status is divorced alive or divorced dead than married. In addition, additional information seen in Components 2 and 3 states that divorce is a scourge that reduces the happiness of Indonesian citizens. In fact, based on BPS data, the divorce rate in 2021 jumped by 53.5%.

In terms of age, Indonesians over 65 years old are relatively happy, and Indonesians who do not feel happy are those between 25-40 years old. In terms of household status, the three components are in line. with spouses feeling happier than heads of households.

In terms of the number of members in the house, when looking at the alignment of the three components. houses with 4 people in the house are more positively perceived as happy.

And in terms of salary, for Component 1. Indonesians feel happy for those who have a salary of less than 1.8 million. On the other hand, there are also Indonesians who feel happier when their salary is more than 7.2 million.

4. Conclusion

GWPCA Gaussian can reduce in knowing the main factors to determine the happiness index of Indonesian citizens in 2021. Based on the results of the analysis. 26 components are formed. but 3 components are sufficient to represent the diversity of the data. which is 83.5%. Based on the implementation of GWPCA Gaussian in the case study of the happiness index of Indonesian citizens. The main factor that determines the happiness of Indonesian citizens related to marital status is being married with 4 family members and this is felt more by wife.

On the other hand, their marital status of divorced alive or divorced dead provides happiness to them. In addition. Indonesians are relatively happier at the age of more than 65 years. Indonesians tend to be positively aligned to feel happy when they have the last education is Diploma I to Diploma III. In terms of salary, Indonesians feel happy when their salary is less than 1.8 million, but many Indonesians also feel happy when their salary is more than 7.2 million.

References

Al. Angela. (2019). Analisis Indeks Kebahagiaan di Indonesia. *Jurnal Jurusan Ilmu Ekonomi*: Vol 7. No 1. Tahun 2019

Anselin. L. (1988). *Spatial Econometrics: Methods and Models*. London: Kluwer Academic Publishers.

Ayundanisa. S.R. dan Sirodj. D.A.N. (2021). Geographically Weighted Principal Component Analysis pada Data Sumber Pendapatan Daerah Kabupaten dan Kota di Jawa Barat Tahun 2019. *Prosiding Statistika*: Volume 7. No.2. Tahun 2021

BPS. (2021). *Indeks Kebahagiaan 2021*. Badan Pusat Statistik RI hal. 104-137

Comber. A.J.. Harris. P. Tsutsumida. N. (2016). Improving land cover classification using input variables derived from a geographically weighted principal components analysis. *ISPRS Journal of Photogrammetry and Remote Sensing*: Volume 119. September 2016. pp. 347-360

Dewi. S.Y. (2020). Determinan Indeks Kebahagiaan di Indonesia. *Jurnal Mahasiswa FEB Universitas Brawijaya*: Vol. 8. No.2

Fotheringham. A.S. Brunsdon. C.. and Charlton. M.. (2002). *Geographically Weighted Regression.* UK: John Wiley & Sons

Harris. P.. Brunsdon. C.. and Charltona. M. (2011). Geographically Weighted Principal Components Analysis. *International Journal of Geographical Information Science* Vol. 25. No. 10. October 2011. pp. 1717–1736.

Harris. P.. Clarke. A.. Juggins. S.. Brunsdon. C.. and Charlton. M. (2015). Enhancements To A Geographically Weighted Principal Components Analysis In The Context Of An Application To An Environmental Data Set. *Geographical Analysis* 47. pp. 146-172

Jolliffe IT. and Cadima J. (2016). Principal component analysis: a review and recent developments. *rsta.royalsocietypublishing.org: Phil. Trans. R. Soc*. A pp: 1-16

LeSage. J. dan Pace. R. Kelley. (2009). *Introduction to Spatial Econometrics*. New York: Taylor & Francis Group.

Mas'ad. Yasin. H.. dan Maruddani. D.A.I. (2016). Analisis Faktor-Faktor Yang Mempengaruhi Persentase Penduduk Miskin Di Jawa Tengah Dengan Metode Geographically Weighted Principal Components Analysis (GWPCA) Adaptive Bandwidth. *Jurnal Gaussian*. Volume 5. Nomor 3. Tahun 2016

Nuroini. H.M. (2019). Perbandingan Jarak Euclidean Dan Jarak Ekonomi Untuk Pemodelan Ekonometrika Spasial Dengan Metode K-Nearest Neighbor (Studi Kasus pada Data PDRB Sektor Pertanian Kabupaten/Kota di Jawa Timur Tahun 2015). Skripsi: Universitas Brawijaya Malang

Pebesma. E. and Bivand. R. (2023). *Spatial Data Science With Applications in R*. New York: Routledge Taylor & Francis Group

Qur'ani. A.Y. (2014). Pemodelan Geographically Weighted Regression Panel (GWR-Panel) Sebagai Pendekatan Model Geographically Weighted Regression (GWR) dengan Menggunakan Fixed Effect Model Time Trend. *Jurnal Mahasiswa Statistik:* Vol. 2 No. 3 (2014) pp. 181-184

Safitri. U. dan Amaliana. L. (2021). Model Geographically Weighted Regression dengan Fungsi Pembobot Adaptive dan Fixed Kernel pada Kasus Kematian Ibu di Jawa Timur. *Jurnal Statistika dan Aplikasinya*: Volume 5 Issue 2. December 2021. pp. 208-220.