SOCIOECONOMIC STATUS FROM SPACE: EXAMPLE OF ESTIMATING THAILAND's SUB-DISTRICT HOUSEHOLD INCOME BASED ON REMOTELY SENSED AND GEOSPATIAL DATA

S. Hutasavi[1], D. Chen[1]

[1] Laboratory of Geographic Information and Spatial Analysis (LaGISA), Dept. of Geography and Planning, Queen's University, Kingston, Ontario, Canada– (16sh48, chendm) @queensu.ca

KEY WORDS: Missing household income imputation, Spatial analysis, Night light intensity, K-NN imputation method, Socioeconomic proxy indicators,

ABSTRACT:

The socioeconomic data, such as household income, is an important indicator of people's well-being. However, due to the limited resource in many developing countries such as Thailand, the data obtained from household income surveys are often incomplete. As a result, the annual household survey usually contains a gap at the municipality household level. In this study, we aim to quantify the household income with K-NN imputation models at the sub-district level using satellite imageries and geospatial data as proxies to socioeconomic indicators. We examined the role of satellite and geospatial data in household income estimation, applied the K-NN imputation methods to estimate the missing income data by using various geographical and statistical variables, and quantified how these data improved the accuracy of sub-district household income estimation. Our results illustrated a significant correlation between sub-district household income and geographical data extracted from day-night satellite data, such as night light intensity (r = 0.53), urban density (r = 0.44), residential area (r = 0.68), urban area (r = 0.64), and statistical data as well as household expenditure (r = 0.97). These can be used to improve the socioeconomic indicators' estimation as well as household income in sub-district level. The income imputation from geographical data perform better result than purely statistical variables. Especially, the night light intensity can infer the wealth of people living in large scale areas, while day-time satellite images can be interpreted for land use and land cover also implying socioeconomic status. Such socioeconomic proxy from space provides spatially explicit information in further study.

1. INTRODUCTION

Since 1992, the sustainable development concepts had been adopted by more than 178 countries. Stakeholders in all counties, particularly the policymakers within lower to middle-income countries, are seriously challenged by emerging sustainable development policies, strategies, plans, and their implementations. The development of Thailand's mega project, the Eastern Economic Corridor (EEC), is framed under the Sustainable Development Goals (SDGs) and required intensive socioeconomic information for the decision making of high-level policymakers.

The socioeconomic data such as household income is an important indicator of people's well-being to eliminate poverty (SDG-1). However, obtaining household income in developing countries, such as Thailand, has been difficult due to inadequate budgeting and time. Although the annual household survey has been done in most areas, the survey does not always cover the municipal area. This problem often leads to a lack of understanding of the economic and sociological standing of people who live in outbound areas. Consequently, this can lead to a lack of area development, economic growth, and limited facilities. Moreover, the limited socioeconomic data can cause difficulty in implementing the national polices as well as disinteresting the vendors to stimulate economic growth.

The EEC area which covers three rural provinces (Rayong, Chonburi, and Chachoengsao) is aimed to improve economic development. Nevertheless, to improve the people's quality of life and well-being, the implementation of EEC requires intensive socioeconomic information for the decision making of high-level policymakers. On the other hand, the process of acquiring statistical information such as economic activities and socioeconomic status can be prohibitively expensive. The lack of surveyed data becomes an important obstacle in developing the economic policies and growth plan in EEC.

Proxy indicators extracted from remotely sensed and geospatial data can be an effective low-cost alternative to an intensive ground survey in low-income countries (Watmough *et al.*, 2019). The proxy indicators have been used to estimate household income by constructing the prediction models (Benin and Randriamamonjy, 2008). Many pieces of research also incorporate remote sensing, geospatial, and statistic data to identify socioeconomic indicators such as poverty (Blumenstock, 2016; Jean *et al.*, 2016; Watmough *et al.*, 2019). Indeed, the night light data is useful for identifying spatially explicit economic activities and socioeconomic status (Doll, Muller and Elvidge, 2000; Elvidge *et al.*, 2009; Bennett and Smith, 2017; Engstrom, Hersh and Newhouse, 2017; Proville, Zavala-Araiza and Wagner, 2017; Dorji *et al.*, 2019). The remotely sensed and geospatial data can also be used to improve and monitor SDGs (Watmough *et al.*, 2019). Additionally, Heitmann & Buri (2019) have shown that remote sensing and geospatial boosting enhanced accuracy and improved traditional household survey methods.

Regression analysis, both linear and nonlinear, is the common technique that has been used to estimate the unknown income value given the observed information. For example, Dai et al., (2012) applied Ordinary Least Squares (OLS) and other regression matrices to construct an income prediction model with 22 related variables. Their result showed the prediction accuracy

at 90% in identifying poor and not poor households in Indonesia. Kalogirou & Hatzichristos (2007) provided an income estimation of households in the municipality of Athens in 2001, capturing a strong non-stationary relationship of education and income across the postal code area.

Motivated by these works, this study aims to estimate the sub-district household income by using the statistical, geographical, and the information extracted from satellite data, replacing the traditional survey. Our study focuses on the socioeconomic status of people in the EEC development area in Thailand. Specifically, we examined the role of satellite and geospatial data and how they improve an income imputation at the sub-district level.

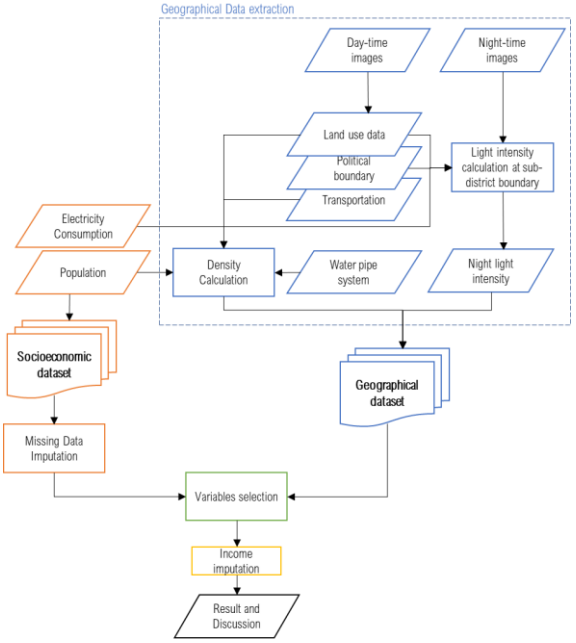

Figure 1. Research flowchart

The research flowchart is shown in Figure 1. This study uses 2016 dataset to estimate sub-district income using the spatial statistics method. Section 2 describes how to generate income proxy indicators and estimation methods. Also, we illustrate the relationship between factors and how to select key indicators for estimating sub-district income. Then, we discuss the result of selecting combination factors and their relationship to sub-district household income in Section 3. Section 4 concludes the role of the geographical data in household income at the sub-district level in Thailand.

2. DATA AND METHODOLOGY

2.1 Data Acquisition and pre-process

In this study, we obtained the sub-district household income from the basic minimum needs survey conducted by the Community Development Department (CDD). Statistical, geographical, and satellite-derived predictors from 2016, was collected from the National Statistical Office (NSO), Department of Public Administration (DOPA), Provincial Electricity Authority (PEA) and Geoinformatics and Space Development Agency (GISTDA), Thailand. The annual night-time light composite from the Visible Infrared Imaging Radiometer Suite (VIIRS) sensor was downloaded via the National Oceanic and Atmospheric Administration (NOAA) website. The Statistical and spatial statistical methods were applied for data preparation and during pre-analysis. The summary of data is shown in Table 1.

Data	Source
Satellite images (daytime)	GISTDA
Land use	GISTDA
Satellite images (night-time)	NOAA
Household income and other socioeconomic data	CDD
Demographic data	NSO
Electricity Consumption by Sector	PEA

Table 1: Source of Statistical and Geographical data used

2.1.1 Land use data (Residential, Commercial, Total Urban area)

In this study, the land use map classified from satellite images was provided by GISTDA. The 2016 land-use dataset was updated using virtual interpretation and ground survey method from the 2011 land-use dataset using THEOS (2 meters panchromatic and 15 meters multispectral resolution) and Landsat 8 images. The misclassification rate of GISTDA's land use data is $\leq 25\%$ at kappa coefficient $\geq 75\%$ and Root Mean Square Error ($RMSE_H$) ≤ 14.4.

2.1.2 Repeated Flood Area

We downloaded the flood data from http://flood.gistda.or.th/ and obtained the frequent flood map from 2005 to 2016 from GISTDA. The flood map was originally extracted from RADARSAT, Landsat, and THEOS during the yearly flooded season. We cropped the data by EEC area and calculated the flood area (km^2) in each sub-district. We also determined the repeated flood area density for each sub-district (km^2).

2.1.3 Population density

We calculated the population density by gathering yearly population data from NSO and DOPA at the sub-district level. Then we calculated the area (km^2) of each sub-district using a political boundary dataset from DOPA. Finally, dividing the population for each sub-district by area, we obtained the population density in each sub-district.

2.1.4 Road density and Waterpipe density

The GIS data of transportation and pipe network from GISTDA was divided into each sub-district in EEC, we calculated the length and summarized the total length in each sub-district area (unit = km.). We used the sub-district area from Section 2.1.3 in sq.km. to calculate the road and pipe density for each sub-district in EEC.

2.1.5 Near Distance

In this experiment, we determined the distance from important places that may influence the household income such as highway, and the capital city (Bangkok).

2.1.6 Night-time light intensity extraction and validation

The night-time light is highly correlated with various socioeconomic indicators such as an area of light, electricity consumption, and population (Proville, Zavala-Araiza and

Wagner, 2017). We studied the Visible Infrared Imaging Radiometer Suite (VIIRS) sensor dataset from NOAA which has been released since May 2012. The dataset consists of monthly and yearly composite images. For this research, we downloaded the yearly composite of a cloud-free dataset for 2016. Hence, to extract the night light intensity value at the sub-district level, we cropped the VIIRS data by study area and calculated by Light intensity using equation 1.;

$$\mu \ ln\, ln\, (L_n) = \frac{\sum_j ln ln\, (I_j)}{A} \quad (1)$$

Where I_j is the intensity for pixel j,
A is total pixels in the sub-district region,
$\mu \ ln\, ln\, (L_n)$ is the light intensity for sub-district n (Nischal et al., 2015).

In our work, we also obtain the daytime satellite images and land use data to help validate nighttime light intensity extraction. It is found that the daytime light intensity has strong correlation with the nighttime light intensity. For example, nighttime light intensities are correlated with daytime satellite images (Jean et al., 2016). Also, Tan (2016) showed that the lit areas had a significant linear relationship with the urban areas at correlation ≥ 0.95.

The satellite images from SPOT 6 and SPOT 7 with the resolution of 1.5 meters were used to quantify the urban density of 48 sub-districts (20% of sub-district in EEC) and compare with the extracted light intensity values. The result is presented in Section 3.1.

2.2 Data Analysis and Variables Selection method

In this study, we examined the role of satellite imageries and geospatial data as proxies to socioeconomic indicators in household income imputation models. The dataset from 2016 was used to examine the role of geographical data in household income prediction at the sub-district level. The 2016 dataset contained 243 records and 22 variables, with 35 records missing of total household income. As the missing records are 14.4%, we then split 10% of complete records for validation of the imputation model and comparison of the performance of predictors.

We argue that the household income at sub-district level could exhibit certain level dependence. The spatial dependence can arise when the observations are collected from points or regions located in space (Lesage and Pace, 2008); therefore, the distribution of the observations will no longer satisfy the property of normal distribution.

The spatial dependence can be represented by the spatial autocorrelation in (2). With the captured spatial dependence, we cannot employ general regression techniques to estimate the sub-district household income.

Moran's I;

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

where n = number of household income in the original matrix
w_{ij} = the spatial weight (W) of general cross-product statistic of household income at location i and j
y = the household income
$\bar{y}$ = the mean of household income

Furthermore, to understand the relationship between predictive (sub-district household income) and predictors , i.e. household incomes and satellite imageries and geospatial data, we determined the relationship among those various variables using Pearson's correlation coefficient (r) to measure the strength of their relationship. The value of r is ranged from -1 (negative relationship) to 1 (positive relationship). A correlation of 0 means there is no relationship between the two variables. Let x and y denote the predictive and predictor variables. The Pearson's correlation formula present below;

$$r = \frac{n(\sum xy)-(\sum x)(\sum y)}{\sqrt{[n\sum x^2-(\sum x)^2][n\sum y^2-(\sum y)^2]}} \quad (3)$$

where; r = Pearson coefficient
n = number of the pairs of variables
$\sum xy$ = sum of products of the pair variables
$\sum x$ = sum of x scores and $\sum y$ = sum of y scores
$\sum x^2$ = sum of squared x scores
$\sum y^2$ = sum of squared y scores

We also checked for multicollinearity using the Variance Inflation Factor (VIF). VIF measures the multicollinearity among variables in the multiple regression model. The large number of VIF indicated a highly collinear relationship to the other variables which should be considered when constructing the prediction model (Table 3).

$$VIF = \frac{1}{1-R_i^2} \quad (4)$$

where R_i^2 = squared of correlation.

The VIF values range from 1, exceeding 4.0, then there is a problem with multicollinearity (Hair et al., 2010). Multicollinearity occurs when two or more predictors in the model are correlated and provide redundant information about the response.

Following the result of r and VIF calculations, we group the predictor variables into four different sets including statistical dataset, geographical dataset, hybrid I dataset, and Hybrid II dataset. The details of each datasets presented in Table 3. The purpose of this process is (1) to determine the performance of statistical and geographical variables, and (2) to examine the role of remote sensing and geospatial data in the household income imputation methods.

2.3 Imputation of missing income data

The household survey often has a gap in the municipality area. Because of this reason, we applied the statistical model to impute the missing value from the available data. Multiple Imputation (MI) is adopted in this research. MI is the best prediction model for yearly income (Ryder *et al.*, 2011). MI gave the distribution of household income which results in more smooth effect and response to middle-income groups as well as Thailand income category (Berzofsky *et al.*, 2015). We applied MI with Rapid Miner Software, using K-Nearest Neighbor (K-NN) operator to impute the missing value.

The K-NN imputation method is suitable for mixed types of variables (Liao *et al.*, 2014). The statistical correlation measurement between different data types has been applied to quantify the Euclidean distance among variables. We then construct the correlation matrix from the classical Pearson

correlation as in Figure 2. This correlation was compared and used to select the K nearest neighbors. The linear regression method is constructed for each neighbor to impute the missing values. Then, the missing household income records were imputed based on each set of predictors (neighbors) following Section 2.2. The outcome was presented in Section 3.3 (Figure 8.).

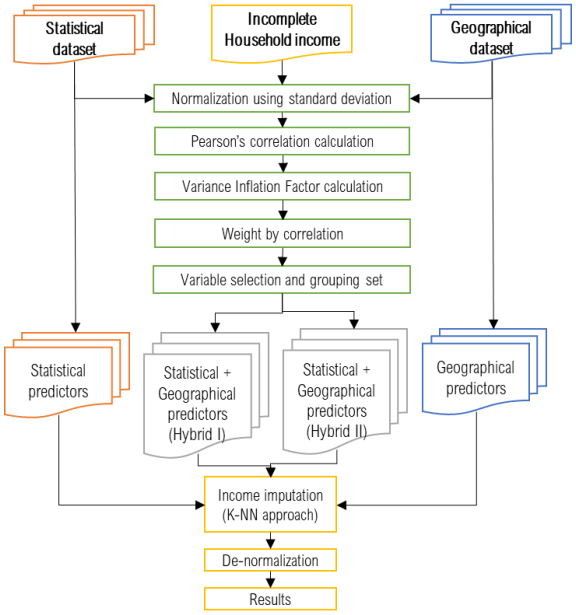

Figure 2. Household income imputation flowchart

2.4 Comparison of the performance of household income predictors

Mean Absolute Error (MAE) was used to summarize and assess the quality of imputation models which generate based on different sets of predictors. MEA is one of model evaluation metrics used with regression models (Sammut and Webb, 2010). It measures the average of errors in each set of predictions by averaging an absolute of difference between prediction and actual value. The MEA equation is shown as below.

$$MAE = \frac{\sum_{i=1}^{n}|y_i - \hat{y}_i|}{n} \quad (5)$$

where $\sum_{i=1}^{n}|y_i - \hat{y}_i|$ = sum of absolute differences between prediction and actual observation, while n = number of samples.

3. RESULTS AND DISCUSSION

3.1 Light intensity map

To extract the night light intensity value at the sub-district level, we cropped the VIIRS data by study area and calculated by equation (1) at the sub-district boundary (Figure 4). We then compared the light intensity with the source data of light as well as urban density in each sub-district which illustrated in Figure 3 and 4. We found that the sum of light intensity correlates to urban density, water pipe network density, and population density at r = 0.80, 0.69, and 0.66, respectively.

We validated the light intensity extraction by comparing it with the urban density. The scatter plot of light intensity and urban density for 48 random sub-districts presented in Figure 5. The lit areas provided a strongly linear relationship with urban density at r = 0.96. Referring to previous research (Tan, 2016), it proved that accuracy of light intensity extraction process is acceptable.

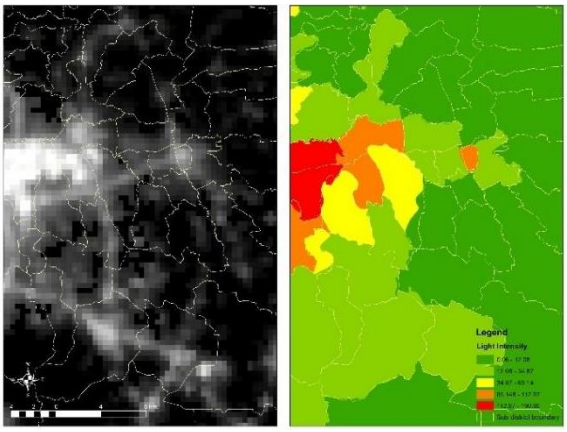

Figure 3. 2016 NTL yearly composite (left) and Night light intensity extraction by a sub-district boundary (right).

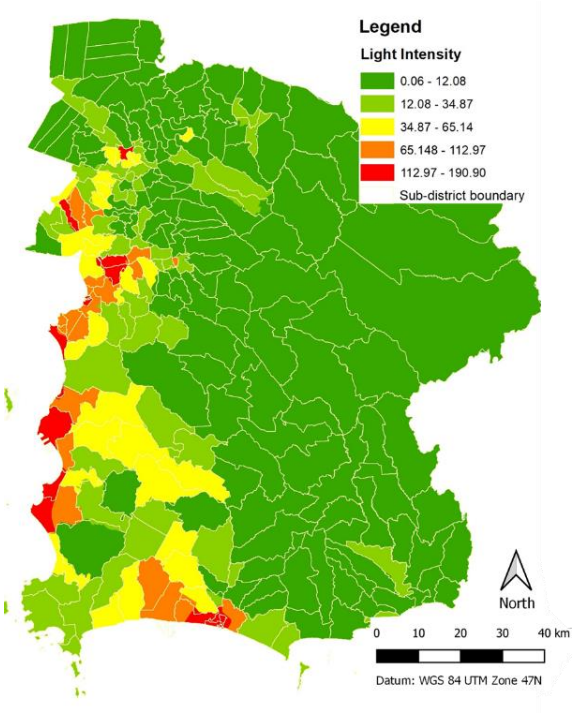

Figure 4. The 2016 light intensity extraction in EEC, Thailand at the sub-district level.

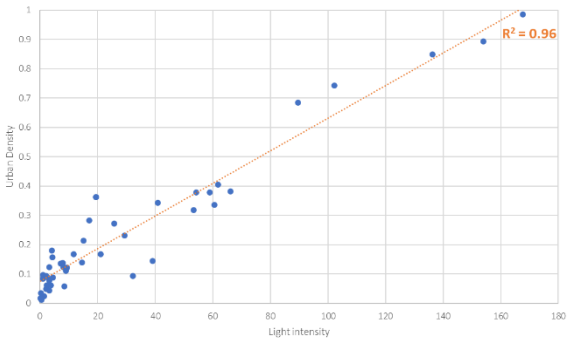

Figure 5. The correlation between night-time light intensity and urban density of 48 sampling sub-districts.

3.2 Data Analytics and Modelling

3.2.1 Sub-district household income descriptive statistic

The general and spatial statistical methods were used to determine the descriptive statistic of the sub-district household income. The result presents an asymmetric distribution which had a long tail to the right-side (Figure 6). Such right-skewed distribution happens when the mass of households is found clustered toward the bottom of the distribution (Donovan, 2015). The group of relatively high incomes at the top draw up the means that make it go beyond the median of household income.

Additionally, the spatial dependence of the total household income data has been calculated by equation (2). The spatial autocorrelation (Global Moran's I) of household income data is 0.62. The z-score is less than 1% likelihood of a clustered pattern. Figure 6 (right) presented that the household income is clustered and has strongly spatial dependence effect. The Moran's I of 0.62 indicated that spatial variables tend to provide a high effect on the sub-district household income. Hence, we prepared the imputation variables from the household yearly survey and geospatial extraction for the imputation model in Section 3.2.2 and 3.2.3.

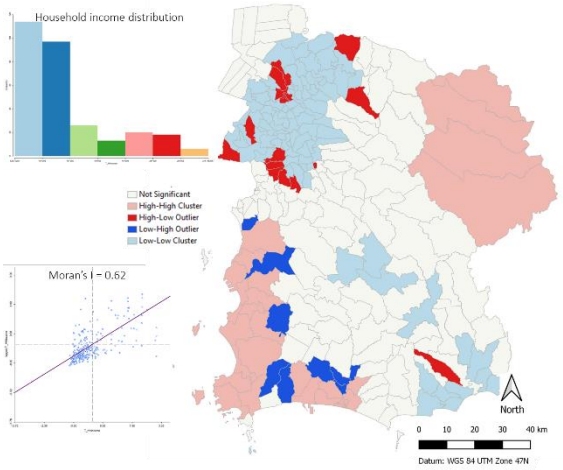

Figure 6. The distribution, spatial autocorrelation (Moran's I), and clustering map of sub-district household income in 2016.

3.2.2 Model variables selection

The statistical analysis of the correlation coefficient between 22 predictors and total household income is illustrated in Figure 7. The correlation between sub-district household income and geographical proxy data, including a residential area, urban area, and sum of light intensity is 0.68, 0.64, and 0.53. These correlation coefficients are less than that of statistical data i.e. total household expenditure, the number of populations, and household (r = 0.63 – 0.97).

Table 2 presented the VIF of each variable. In this study, we assign the variables that give VIF more than 10 as the multicollinearity variables. The total population has a very strong multicollinear with a sub-district household income at VIF = 951.23.

From the calculation of r and VIF, the house expenditure is the best variable which provides highest r at 0.97 and low VIF at 5.96. For the geographical data, commercial area and night-time light intensity were significant variables for the household income estimation model. They provided the r value more than 0.5 and VIF lower than 5.0. Especially, the night-time light intensity provided medium correlates to sub-district household income with lower VIF (r = 0.53 and VIF = 2.29), see more information about r and VIF values in Table 2.

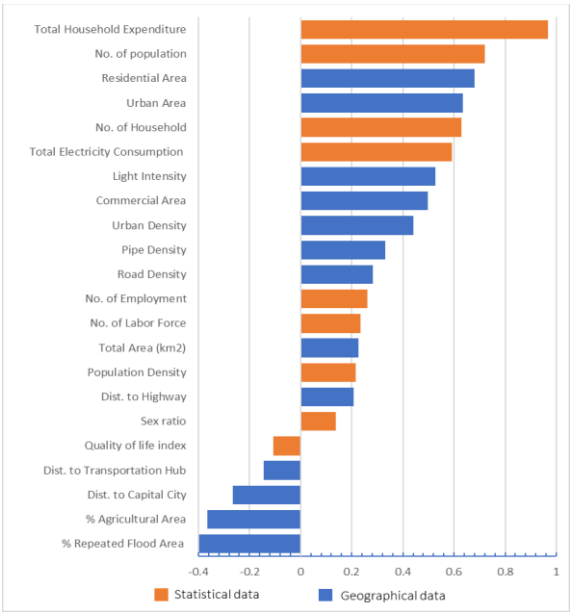

Figure 7. The correlation coefficient between predictors and sub-district total income.

Based on the result of correlation and VIF calculation, we separated variables into four groups of predictors including, statistical, geographical, hybrid I, and hybrid II. The statistical[1] and geographical[2] dataset comprises all variables in each category. The hybrid dataset combines both types of the dataset, hybrid I composes of six variables which give $r \geq |0.4|$ and $VIF \leq 10$. The hybrid II dataset includes nine variables that provide r more than 0.4 and ignoring VIF value. The details of each dataset are shown in Table 2.

	Statistical[1]	Geographical[2]	Hybrid I[3]	Hybrid II[4]	r	VIF
Quality of life index	●				-0.11	1.40
Sex ratio	●				0.14	4.73
No. of Labor Force	●				0.23	44.04
No. of Employment	●				0.26	44.18
Total Electricity Consumption	●		●	●	0.59	16.92
No. of Household	●			●	0.63	25.41
No. of population	●			●	0.72	951.23
Total Household Expenditure*	●			●	0.97	5.96
% Repeated Flood Area		●	●		-0.40	5.81
% Agricultural Area		●	●		-0.37	4.59
Dist. to Capital City		●			-0.26	3.76
Dist. to Transportation Hub		●			-0.14	26.57
Dist. to Highway		●			0.21	2.21
Population Density		●			0.22	5.81
Total Area (km2)		●			0.23	11.15

	Statistical[1]	Geographical[2]	Hybrid I[3]	Hybrid II[4]	r	VIF
Road Density		●			0.28	1.55
Pipe Density		●	●		0.33	4.85
Urban Density		●		●	0.44	10.17
Commercial Area		●	●	●	0.50	4.68
Light Intensity*		●	●	●	0.53	2.29
Urban Area		●		●	0.64	16.99
Residential Area		●		●	0.68	21.57

Table 2. The different four sets of predictors with Pearson's correlation (r) and Variance Inflation Factor (VIF) values.

3.3 Income imputation model implementation

We then applied the imputation model using the K-NN function in RapidMiner software to estimate the missing household income values. Firstly, we used all variables (both statistical and geographical) to impute the original missing data (35 records). We assumed this result as a reference household income data in Figure 8 (3). Secondly, we randomly remove 10% more from original data (25 records) which is shown as the training data in Figure 8 (2). Thirdly, we imputed the 60 missing household income data using the four different datasets from Section 3.1.2. The imputation results from each difference set of variables were illustrated in Figure 8 (4), (5), (6), and (7).

To answer the research question --- how satellite and geospatial data improved the sub-district household income imputation model; (1) we investigated the variables selection method in Section 3.2. We found that the geographical data was a significant variable for the estimation model as present in Table 2. (2) In this section, we examined the role of remotely sensed images and geospatial data by testing the four different sets of variables in K-NN imputation method. Then, we compare their performance using MAE. The overall performance of each set of predictors were shown in Table 3.

Predictors set	MAE
Statistical	50094238.98
Geographical*	**34151656.23**
Hybrid I	40078309.51
Hybrid II	35296612.54

Table 3. The mean absolute error of imputation models (K-NN approach) based on sets of predictors.

The set of geographical variables were the best predictors for imputing household income using K-NN function. They provided the lowest MAE. This result proved that there is a strong spatial effect in household income at sub-district level. Also, it suggested that the household income imputation model with the geographical variables performs much better than the model that formulates from purely statistical and hybrid variables.

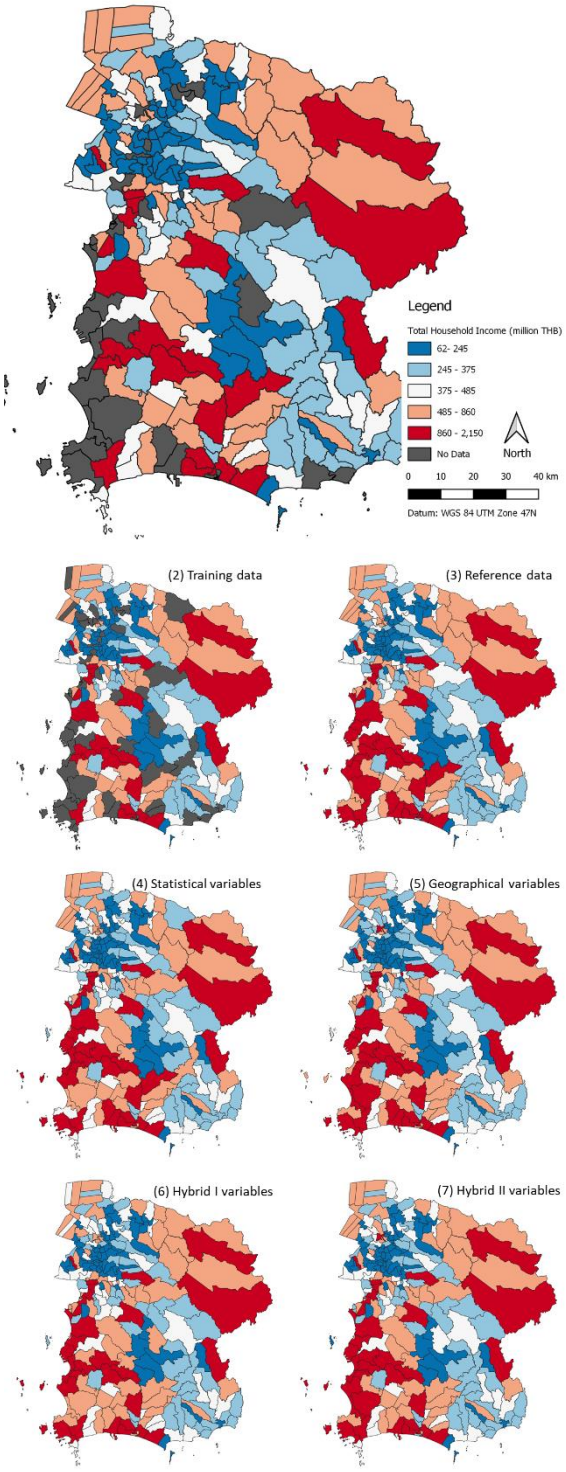

Figure 8. The imputation results of missing sub-district household income by four different sets of predictors.

For the next step, we will improve our household income estimation model using the time – series data set from 2011 to 2019. Also, we will integrate night light intensity as a proxy of wealth of people and human activities, as well as utilize day-time satellite images to identify settlement characteristics. Such data can be used to infer socioeconomic status from space.

4. CONCLUSION

The incorporation of remotely sensed, geospatial, and statistical data improves the accuracy of the estimation of total household income at the sub-district level. The night light intensity, land-use area, land use density, were useful to estimate household income at the sub-district level. The improvement from incorporating geospatial and satellite-derived datasets illustrated the potential to bridge the gap from a traditional socio-economic survey in Thailand, especially on household income at the municipal level. This study proves that the night light intensity can be used as a proxy of the wealth of people and human activities in sub-district level. Such data can be used to infer socioeconomic status from space. High-level policy planning, including the EEC development plan, can benefit from a more complete and spatially explicit insight on the socio-economic situation.

ACKNOWLEDGEMENT

We would like to thank the National Statistical Office (NSO), Department of Public Administration (DOPA), Community Development Department (CDD), and Geoinformatics and Space Development Agency (GISTDA), Thailand for sources of statistical and geographical data. Additionally, we also highly appreciate the help from Nuntikorn Kitratporn and Suwichaya Suwanwimolkul, who dedicated their time to proofread the manuscript.

REFERENCES

Benin, S. and Randriamamonjy, J. (2008) 'Estimating Household Income to Monitor and Evaluate Public Investment Programs in Sub-Saharan Africa', (June), p. 32.

Bennett, M. M. and Smith, L. C. (2017) 'Advances in using multitemporal night-time lights satellite imagery to detect, estimate, and monitor socioeconomic dynamics', *Remote Sensing of Environment*. doi: 10.1016/j.rse.2017.01.005.

Berzofsky, M. *et al.* (2015) 'Imputing NCVS Income Data'.
Blumenstock, J. E. (2016) 'Fighting poverty with data', *Science*, 353(6301), pp. 753–754. doi: 10.1126/science.aah5217.

Dai, J., Sperlich, S. and Zucchini, W. (2012) 'Estimating and Predicting Household Expenditures and Income Distributions', *SSRN Electronic Journal*, 41(0), pp. 1–26. doi: 10.2139/ssrn.1965409.

Doll, C. N. H., Muller, J. P. and Elvidge, C. D. (2000) 'Night-time imagery as a tool for global mapping of socioeconomic parameters and greenhouse gas emissions', *Ambio*, 29(3), pp. 157–162. doi: 10.1579/0044-7447-29.3.157.

Donovan, S. A. (2015) 'A guide to describing the income distribution', *Income Distribution in the United States: Measures, Trends and Analyses*, pp. 1–39.

Dorji, U. J. *et al.* (2019) 'A machine learning approach to estimate median income levels of sub-districts in Thailand using satellite and geospatial data', pp. 11–14. doi: 10.1145/3356471.3365230.

Elvidge, C. D. *et al.* (2009) 'A global poverty map derived from satellite data', *Computers and Geosciences*, 35(8), pp. 1652–1660. doi: 10.1016/j.cageo.2009.01.009.

Engstrom, R., Hersh, J. and Newhouse, D. (2017) 'Poverty from Space: Using High-Resolution Satellite Imagery for Estimating Economic Well-Being', (December), p. 36. Available at: http://econ.worldbank.org.

Heitmann, B. S. and Buri, S. (2019) *Poverty Estimation with Satellite Imagery at Neighborhood Levels, International Finance Coperation*. Available at: https://www.ifc.org/wps/wcm/connect/2cae89ee-dea3-4a7e-ba79-77c9011cbd0f/IFC_2019_Poverty+Estimation+with+Satellite+Imagery+at+Neighborhood+Levels.pdf?MOD=AJPERES&CVID=mHZhcxB.

Jean, N. *et al.* (2016) 'Combining satellite imagery and machine learning to predict poverty', *Science*. doi: 10.1126/science.aaf7894.

Jean, N., Luo, R. and Kim, J. H. (2016) 'Nighttime Light Predictions from Satellite Imagery'. Available at: http://cs231n.stanford.edu/reports/2016/pdfs/423_Report.pdf.
Kalogirou, S. and Hatzichristos, T. (2007) 'A spatial modelling framework for income estimation', *Spatial Economic Analysis*, 2(3), pp. 297–316. doi: 10.1080/17421770701576921.

Lesage, J. P. and Pace, R. K. (2008) 'Spatial Econometric Modeling Of Origin-Destination Flow ∗', *Journal of Regional Science*, 48(5), pp. 941–967. doi: 10.1111/j.1467-9787.2008.00573.x.

Liao, S. G. *et al.* (2014) 'Missing value imputation in high-dimensional phenomic data: Imputable or not, and how?', *BMC Bioinformatics*, 15(1), pp. 1–12. doi: 10.1186/s12859-014-0346-6.

Nischal, K. N. *et al.* (2015) 'Correlating night-time satellite images with poverty and other census data of India and estimating future trends', *ACM International Conference Proceeding Series*, 18-21-Marc(April 2016), pp. 75–79. doi: 10.1145/2732587.2732597.

Proville, J., Zavala-Araiza, D. and Wagner, G. (2017) 'Night-time lights: A global, long term look at links to socio-economic trends', *PLoS ONE*, 12(3), pp. 1–12. doi: 10.1371/journal.pone.0174610.

Ryder, A. B. *et al.* (2011) 'The advantage of imputation of missing income data to evaluate the association between income and self-reported health status (SRH) in a Mexican American cohort study', *Journal of Immigrant and Minority Health*, 13(6), pp. 1099–1109. doi: 10.1007/s10903-010-9415-8.

Sammut, C. and Webb, G. I. (eds) (2010) 'Mean Absolute Error BT - Encyclopedia of Machine Learning', in. Boston, MA: Springer US, p. 652. doi: 10.1007/978-0-387-30164-8_525.

Tan, M. (2016) 'Use of an inside buffer method to extract the extent of urban areas from DMSP/OLS nighttime light data in North China', *GIScience and Remote Sensing*. Taylor & Francis, 53(4), pp. 444–458. doi: 10.1080/15481603.2016.1148832.

Watmough, G. R. *et al.* (2019) 'Socioecologically informed use of remote sensing data to predict rural household poverty', *Proceedings of the National Academy of Sciences of the United States of America*, 116(4), pp. 1213–1218. doi: 10.1073/pnas.1812969116.