THE UNIVERSITY OF CHICAGO

Patterns and Factors in Aging Population Distribution in Japan

By Dhruval Bhatt

July 31st, 2020

Faculty Advisor: Dr. Luis Bettencourt

Preceptor: Joshua Mausolf

A paper submitted in partial fulfillment of the requirements for the Master of Arts degree in the Master of Arts in Computational Social Science

Abstract

Japan, a leading site of productivity and economic growth, is facing serious challenges due to a rapidly aging population. To understand and address the uneven distribution and concentration of aging population, demographic transition theory is considered. An exploratory spatial analysis using the data published by Japanese government for each of the 47 prefectures (administrative units), reveals that highest concentration of aging population is in less industrialized, rural regions and regions with prominent industries and urban regions have lowest concentration of aging population. A model with direct factors for population, birth rate, death rate and migration, explains ~92% of ratio of population over 65 but is diagnosed with a spatial error, indicating missing variables or correlation in the error terms. On the other hand, a model for indirect factors of population aging has dependent variables, such as, densely inhabited districts characterizing urban areas, fertility rate showing changing social norms as well as income, education, industry and job-related migration, shows presence of a spatial lag. After accounting for that, the model explains ~63% of ratio of population over 65. This is study shows the challenges of working with a limited number of administrative units in spatial regression analysis and provides a foundation for further research in role of urbanization in the distribution of aging population in Japan.

Key words: Demography, Population Aging, Urbanization, Spatial Analysis, Japan

Background and Motivation

The world's third largest economy, Japan has been the leading site of innovation and growth for several decades. However, currently, the Japanese economy is facing challenges such as low growth in productivity, labor force decline, concentration of economic growth in major metropolitan areas and uneven distribution of aged population (OECD, 2016). With statistics that indicate that "that the working age population is falling from 69.5% of total in 1995 to 62.1% in 2013 and further to 50.9% in 2060", the super-aged Japanese society is proactively addressing issues of population decline and demographic shift. (West, 2014). The government is focusing on policies that include participation of women and elderly in the workforce, leveraging technology to improve productivity and services rendered, and harnessing spatial planning and spatial policies to tackle regional disparities.

The focus on spatial variation and policies to address them is an important one. As early as 1960s and 70s, scholars and government had recognized that of the population problems, "the increase of elderly people and the uneven distribution of population are thought to be the most important" (Kawabe, 1980, p. 192) Since then, the problem has only intensified. Japan is an island nation with mainly mountainous terrain that constrains most of the inhabitable area to low lying plains leading to high density cities. For instance, the "Population per 1 km^2 of total land area" in 2010 is 343.4 but "Population per 1 km^2 of inhabitable area" is 1048.4. In 2015, those values are 340.8 and 1036.4. This shows that the population is declining but still the inhabitable areas are very dense. The map below shows the distribution of population per 1km^2 of inhabitable area across prefectures.

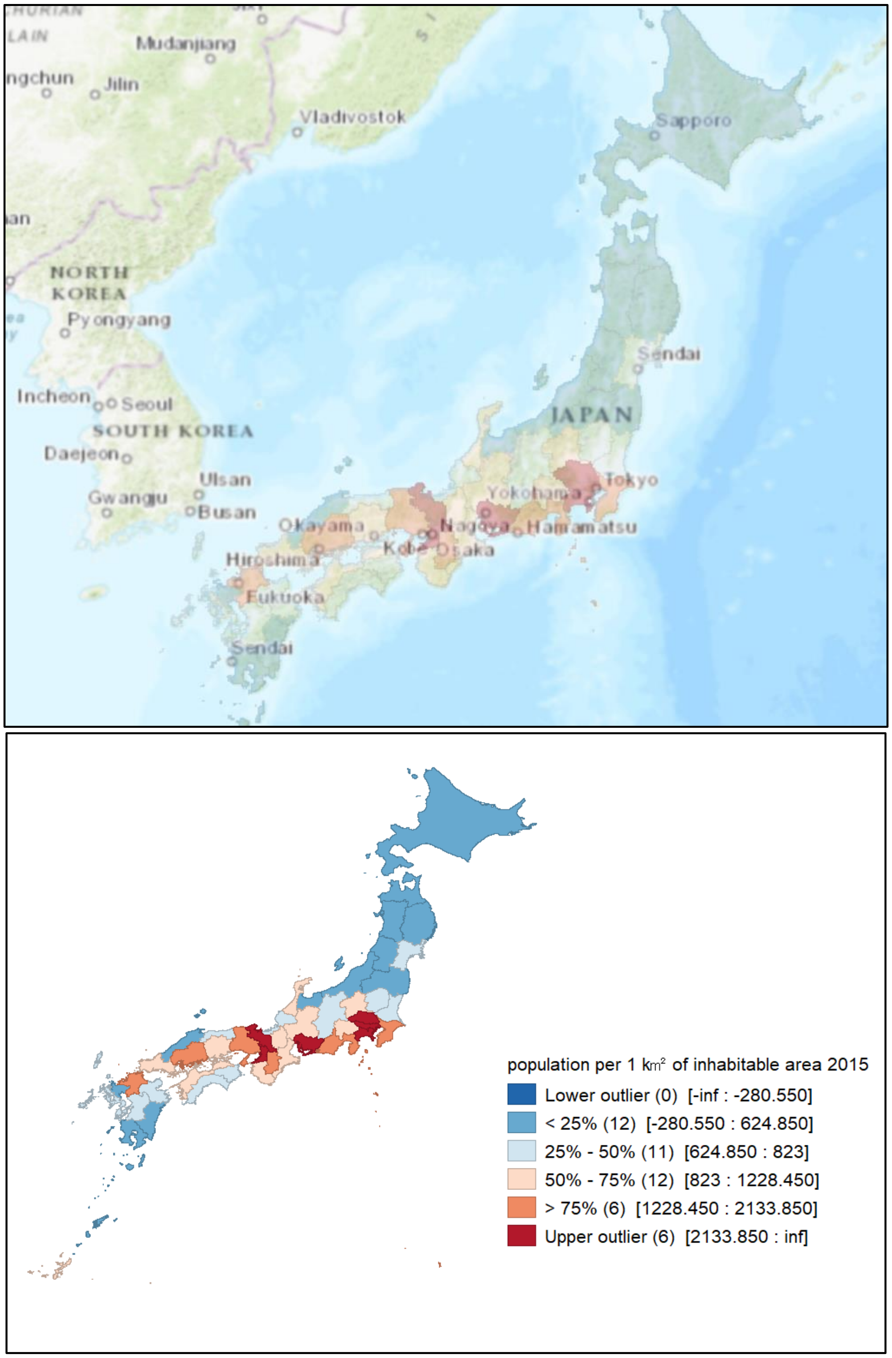


Figure 1 Population Concentration across Inhabitable Lands

From this map, it is evident that regions with major cities like Tokyo, Nagoya, and Osaka are the upper outliers in population concentration while the further out northern and southern regions are less dense. The regions with large concentration of population can leverage economies of scale, have access to human capital and can be highly innovative. However, dense cities also put a burden on the environment in the area, have a high cost of living and lead to depopulation of rural areas. Japanese government's commitment to infrastructure and service provision has ensured that "despite increasing concentration of activity and population, Japan in 2010 recorded the second lowest inter-regional Gini co-efficient for the GDP per capita in the OECD and the lowest disparities in the OECD between predominantly urban and rural regions" (OECD, 2016, 11). However, in the face of changing demographics, this spatial heterogeneity in population concentration becomes a matter of concern.

In addition to the uneven distribution of population, particularly, there is a variation in the distribution of aging population. However, the general pattern is the opposite. As seen in the maps below, unlike population distribution, the high outliers of aging index in 2015 are in the rural and mountainous regions of the country.

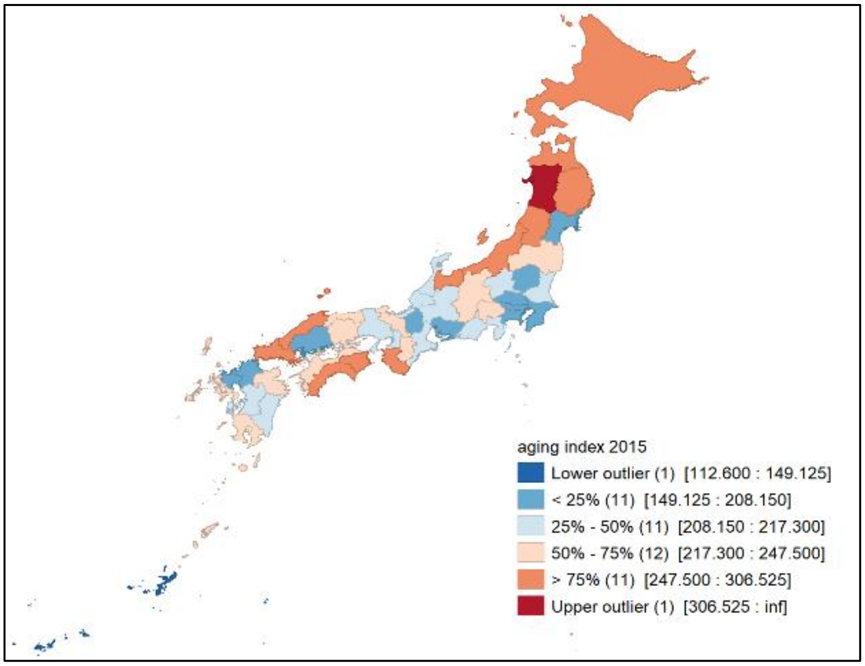


Figure 2 Distribution of Aging Index in 2015

Introduction

Carefully studying the spatial patterns of aging population and understanding characteristics of regions could provide an insight into the relation between uneven distribution of population density due to urbanization and concentration of aging population. This analysis answers questions such as: does exploratory spatial data analysis indicate a concentration of aging population in Japan? Is there spatial dependency in the direct factors to aging population? Are there other predictable characteristics of regions that leads to concentration of aged population? Do those factors have spatial dependence? The goal of this paper is to use spatial analysis to understand the factors for the patterns in concentration of aging population in Japan.

Stable demographic patterns had been a norm for most of time, despite the fluctuation in death rates as birth rates have remained high historically lending stability in population growth (Olshansky, 1997). However, this pattern has been changing in the last century, especially in industrialized, economically developed regions. Theory suggests that population dynamics evolves from the Malthusian model to Post-Modern times with industrialization. Centuries ago, Thomas Malthus projected that population would continue growing exponentially unless it is limited by resources. In the Malthusian model, the population increases but the income growth decreases as the resources are shared among growing population. However, as cities progress, they see the phenomenon of Post-Malthusian Regime where the population continues to grow but rapid increase in technology and total output leads to higher growth rate in income as well. In contrast, the Post-Modern Regime sees a high growth rate for income due to modernization but a lowered population growth (Galor and Weil, 2000; Lagerlof, 2000).

The Post-Modern Regime is characterized by high scientific and technological progress and economic prosperity in that society. This has direct as well as unintended impact on the three tenets of population dynamics – fertility, migration, and mortality. The Demographic Transition Theory is a widely accepted generalization that states that fertility rates decline as societies modernize due to emphasis on quality over quantity of children or reduced importance of offspring and mortality rates decline due to better medical care and disease control (Kirk, 1996). This transition is usually beneficial in the early stages as the resources and rewards per capita is increased but becomes problematic when the proportion of aged population, likely not a part of productive labor force, is much greater than the working age population (Galor, 2011).

Many industrialized nations are currently inching towards the final stages of this transition as population continues to age but birth rate and in-migration is not growing at a fast-enough rate. Scholars have shown that declining mortality rates is a dominant factor in demographic shift to a rapidly ageing population (Preston, Himes, Eggers, 1989). It is imperative to understand the cause, trajectory, and effect of this impending phenomenon in modernized society for long term stability. While many scholars of Demography accept the notion of demographic transition, the concept is ambiguously used as a theory, historical model or a predictive model or is often just descriptive (Szreter, 1993). Therefore, the factors affecting the varying rate of population fertility, mortality, and migration are not consistently predictable and require a deeper understanding in context of the society in consideration.

The “coordinated development of the economy and society, the advancement of science and technology, the development of urbanization, and the institutional changes in the process of modernization are the driving mechanisms of demographic transitions” (Wu, Song and Yu, 14, 2019). In the words of a notable Demographer, Frank Notestein, “the "problem of aging" is no problem at all. It is only the pessimistic way of looking at a great triumph of civilization” (Notestein, 1954). Large urban areas become centers of sustained economic growth and innovation, leading to longer life expectancy, high migration of working age population and an environment for social and culture transformation on views on participation of women in workforce or ideal time and number of children to have. The growth and spread of these ideals can be institutionally driven through structures present in workforce, education or healthcare as well as organically diffused through cultural norms.

The spatial significance of factors affecting aging population is undeniable. In addition to the urban-rural divide being on a spatial scale, characteristics, such as, gross domestic product of the region, education level, type of industry, as well as socio-economic views could have spatial autocorrelation (Kim Parker, 2020). Therefore, classic econometric methods that just consider the economic and environmental factors on population aging is inaccurate as it disregards the “spatial dependence and heterogeneity of population aging among regions” (Wu, Song and Yu, 14, 2019). In the mid twentieth century and beyond, Geographies of Ageing was an emergent field that focused on uneven distribution of aging population and utilized spatial trends to incorporate new aspects of social theory to understand aging (Davies and James, 2016). The inclusion of human geography in gerontology and aging studies was considered a gateway to new ideas in demographic social theory (Harper and Laws, 1995).

“Demography is an inherently spatial science” and increasingly available spatial data, spatial analysis tools and methods aid in the growth of this area of study (Matthews & Parker, 2013). The field of spatial data science is rapidly growing to include space as an important feature in exploration. This paper utilizes the software and methodology that is proposed by Anselin and his team working on GeoDa, a spatial analysis software, to explore the spatial significance of factors contributing to aging population (Anselin, Syabri, and Kho, 2009). Specifically, scholarly examples of spatial methods applied in Japan, include, prefecture level analysis for telecommunications divide (Nishida, Pick, and Sarkar, 2014), Tokyo ward-level analysis for diffusion of burglary (Shimada, 2004) and local spatial correlation of aging population measures (Shiode, Morita, Shiode and Okunuki, 2014). Additionally, research on spatial analysis of aging population includes spatial perspective on living arrangements among elderly in Europe (Szołtysek, Ogórek, Poniat, and Gruber, 2019), exploring spatial variation in population aging in China’s mega cities (Xie, Zhou and Luo, 2015) and impact of social, economic and environmental factors on population aging in China (Xu, Zhao, Zhang, and Xia, 2018).

Scholarly work in spatial analysis of population and demographic trends tends to be heavily focused on individual behavior in a sub-population or limited to inferences from exploratory analysis. However, this paper utilizes prefecture-level (Japanese administrative units) aggregate data for various social indicators to build a spatial regression model. Though this greatly limits the amount of data points available and restricts the population inferences to a higher level, it provides a national overview of regional variation. This paper follows a similar spatial analytical approach as authors of “Spatial Differences in China’s Population Aging and Influencing Factors: The Perspectives of Spatial Dependence and Spatial Heterogeneity” for understanding aging population in Japan (Wu, Song and Yu, 2019). Compared to China, Japan is much smaller with a fraction of administrative units to consider and therefore it constricts the number of exploratory variables that can be included in the model. To guide the method for feature selection, decision tree driven, feature importance charts are generated, and selected variables are evaluated through iterations of spatial regression models. The combination of demographic transition theory, machine learning method and spatial analysis proves to be effective in lending insights to factors and patterns in aging population distribution in Japan.

While the methodology adds to the literature of spatial analysis in demography, the results of this study will serve as essential information for the government to determine appropriate regional policies and allocate resources to provide critical services to the population. The Japanese government already recognizes this and considers spatial policy as an important aspect for national economic well-being. “The new National Spatial

Strategy (NSS) aims to sustain a settlement pattern that facilitates the realization of agglomeration economies while avoiding the abandonment of large parts of the national territory" and encourages a regional focus on unique assets to attract and retain population (OECD, 2016, 12). Doing so requires an in-depth understanding of the impact of urbanization and other factors that affect population at large.

This paper explores the patterns in distribution of aging population in Japan and models the factors contributing to concentration of aged population. It starts with a brief overview of the data, followed by a methods and results section. The methodology starts with exploratory spatial data analysis followed by model specification and spatial regression models with both direct and indirect factors to aging population. The paper ends with a discussion on insights from the results and addressing any limitations encountered.

Data

The primary source of data is government published statistics for 47 prefectures (first level of jurisdiction and administrative division) in Japan, available on e-stat (Statistics of Japan). The data spans 12 different topics with several variables available in each of the topics. The chart below summarizes the available variables across topics.

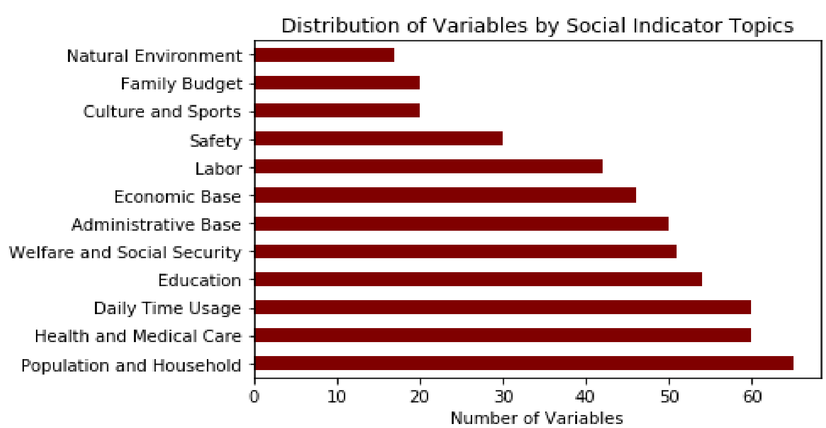


Figure 3 Chart of Topics and Variables Available

Currently, the most complete data is available for year 2010 and 2015 – hence, those are the main years in consideration for analyzing any temporal variation here. Using a combination of manual organization and python scripts, a subset data frame for each year with only variables that are available for both 2010 and 2015 is created for analysis. This is a list of 453 variables in total that characterizes various social, economic, and natural features of the 47 prefectures. For spatial analysis, since the data is available for prefecture level, geopackage files, with layers for national, prefecture boundary and municipalities, is downloaded from GADM maps and data (GADM Japan 2018).

Method and Initial Results

To explore the spatial process in the factors contributing to uneven distribution of aging population across Japan, a multistep methodology is applied. It starts with an exploratory analysis of aging population, spatial regression analysis of direct factors of aging population and finally, variable selection and spatial regression analysis of indirect factors contributing to aging population.

Exploratory Analysis: Spatial Autocorrelation

There are many indicators of aging population, such as, 'ratio of aged population', 'aging index', 'ratio of aged population'. Given that the "aging population ratio can reflect the absolute population aging and is the most frequently used indicator" (Wu, Song and Yu, 5, 2019), for this analysis, ratio of population [65 years old and over] (or the proportion of people over a retirement age compared with the total population) is chosen. The figure below shows a simple choropleth map of the distribution of this variable.

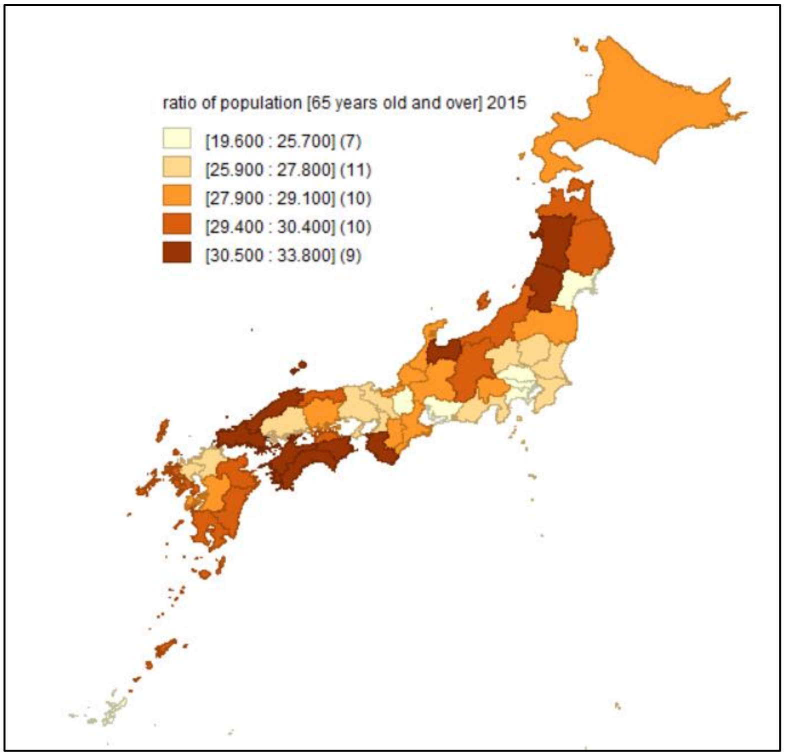


Figure 4 Distribution of Aging Population

This map indicates potential spatial autocorrelation so global and local correlation tests are conducted to explore this further. However, prior to any spatial exploration or regression, a weight matrix that accurately specifies the neighbors to consider is required. A distance based, k-nearest neighbor (k-NN) weights matrix is specified as contiguity-based neighbor specification would not be adequate to capture the numerous island prefectures, that do not share borders with their neighbors. The 'k' in k-NN represents the number of neighbors a prefecture has. Starting from the default setting of 4 neighbors, the neighbors are varied to 5, 6 and 7. Finally, k-NN with 7 neighbors is selected based on the interesting clustering pattern in local autocorrelation and theoretically it is justified as metropolitan areas and rural areas tend to encompass a large area.

The figures below show the results of global Moran's I statistic. "The Moran scatterplot is an illustration of the relationship between the values of the chosen attribute at each location and the average value of the same attribute at neighboring locations" (Blanford et. al, 2018). Most values are in top right and bottom left, showing that there is an overall positive autocorrelation. The high value quadrant is highlighted and represented in the middle picture. The results of randomization on the far-right show that the Moran's I statistic is in fact significant.

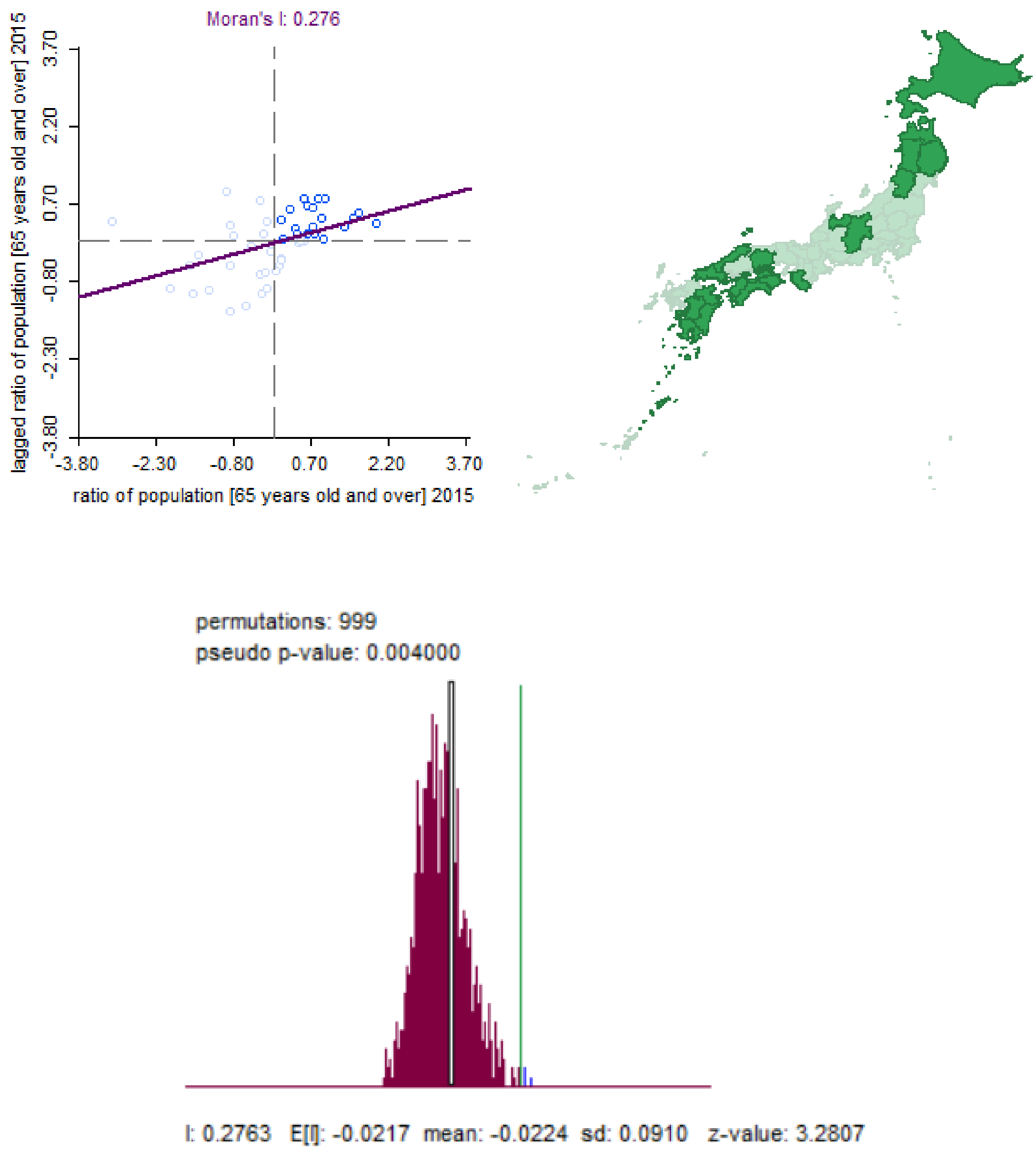


Figure 5 Results of Moran's I Analysis for Global Autocorrelation

While results above confirm there is global autocorrelation, it is not as meaningful as the results could be dominated by clusters of locations with high correlations. To explore an informative pattern of local clusters, Local Indicators of Spatial Autocorrelation (LISA) is conducted on GeoDa. The results below show local clusters for ratio of aging population over all population with both significance level of 0.05 and 0.01 with a seed value set to 0.

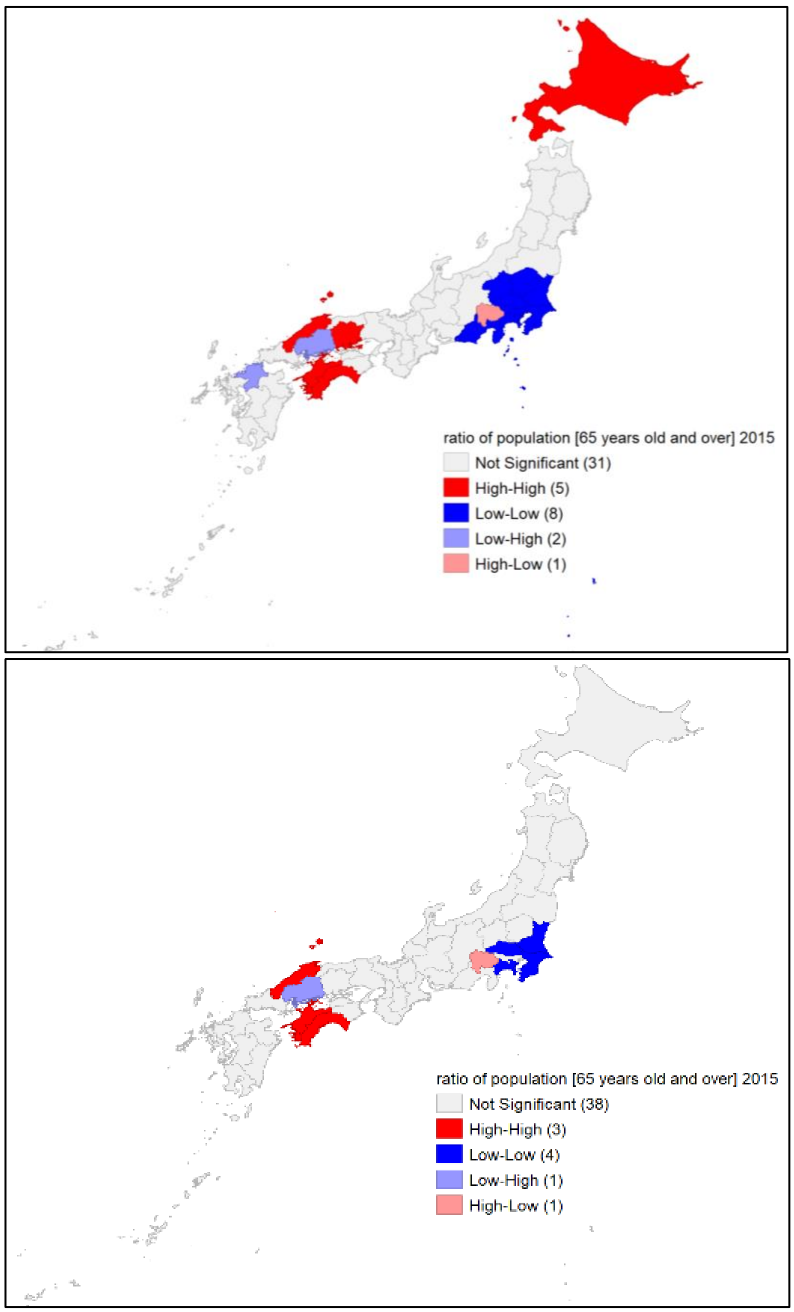


Figure 6 Local Indicators of Spatial Autocorrelation KNN = 7: p < 0.05 and P <0.01; seed = 0

From the results above, it is seen that there are two main areas: one of high-high cluster and one of low-low cluster. Interestingly, both have outliers of high-low pattern and low-high pattern, respectively. The northern most prefecture is not a low-low cluster with a more stringent p-value cutoff. Figure 17 and 18 in the appendix shows LISA results for the same variable using different number of neighbors for the k-NN weights matrix and the pattern for ratio of population [65 years old and over] in 2010. From the results of the 2010 data, it is evident that not much is different in the concentration of aging from 2015. Therefore, temporal analysis is not pursued. To develop an intuition for what may be causing the patterns seen in figure 6, the 0.05 significance level map is mapped on the geographical base first and a reference map of the 11 metropolitan areas in 2015 shown on the bottom half of figure 7.

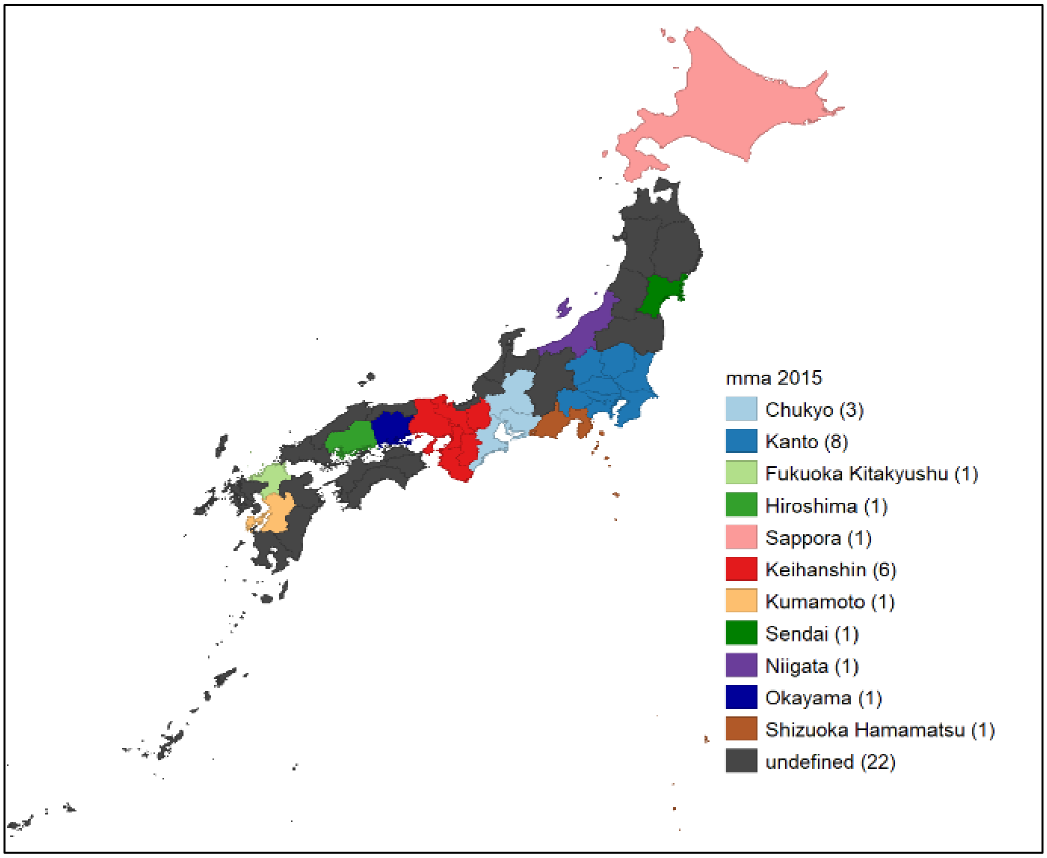


Figure 7 Geographical Map with LISA and Metropolitan Areas Identified; Undefined are not Part of Any

From the LISA analysis, the High-High cluster is composed of Shizuoko, Kanagawa, Chiba, Gunma, Tochigi, Ibaraki, Tokyo, Saitama. All of them, except Shizuoko, are a part of Kanto MMA (Major Metropolitan Area). With the capital city at its center, this group is part of the most populous and wealthiest metropolitan area. Its High-Low outlier, Yamanashi, is mostly mountainous region with mostly agrarian economy with presence of some small-scale industries, mostly in Kōfu city (Pletcher, 2013).

On the other hand, the Low-Low clusters include Hokkaido (in the north) and Shimane, Ehime, Kochi, Okayama (in the south). Of these, only Okayama and Hokkaido have the population to be considered metropolitan areas, and both have significant coverage of national parks and no major industries. Its Low-High outliers include

Hiroshima and Fukuoka. Hiroshima is in a plain and is the center of “the region’s largest conurbation’ and Fukuoka has a port and is central to commercial activities in the region (Ray 2018 and Pletcher 2013). Both are major industrial centers – Hiroshima is the headquarters for automobile company, Mazda and Fukuoka is the birthplace of tire manufacturer, Bridgestone.

From this quick overview, it is evident that factors such as geography, industrialization, population density may have an impact on the concentration of aging population. This lends merit to the theory that impact of urbanization affects the factors that lead to concentration of aging population and given the spatial nature of those processes, any underlying spatial dependency in the clusters should be understood.

Effect of Direct Factors

From the theoretical discussion, it is known that change in birth rate, death rate and migration rate are the principal factors leading to population aging. A spatial regression analysis with just these factors is conducted in GeoDaSpace, a spatial analysis software. The screenshot for the results of the analysis is shown in figure 8 and 9. The results of the regression show that those direct factors, can explain 92% of the variation in ratio of population over the age of 65. Of the explanatory variables, the birth rate (crude birth rate (per 1,000 persons)) and death rate (crude death rate (per 1,000 persons)) is significant but rate of net migration is not. Unsurprisingly, the birth rate coefficient is negative, indicating inverse correlation with concentration of aging population, while the death rate coefficient is positive, indicating the opposite. The diagnostic test results for this regression show that it has some issues with multicollinearity and heteroskedasticity but has normality of errors.

```
REGRESSION
----------
SUMMARY OF OUTPUT: ORDINARY LEAST SQUARES
-----------------------------------------
Data set            :finalmodel_var3.dbf
Weights matrix      :File: SRA_DB_JPN_knn7.gwt
Dependent Variable  :  aged_ratio                Number of Observations:        47
Mean dependent var  :     28.2851                Number of Variables   :         4
S.D. dependent var  :      2.7671                Degrees of Freedom    :        43
R-squared           :      0.9244
Adjusted R-squared  :      0.9191
Sum squared residual:      26.628                F-statistic           :   175.2623
Sigma-square        :       0.619                Prob(F-statistic)     :  3.963e-24
S.E. of regression  :       0.787                Log likelihood        :    -53.337
Sigma-square ML     :       0.567                Akaike info criterion :    114.675
S.E of regression ML:      0.7527                Schwarz criterion     :    122.075

------------------------------------------------------------------------------------
            Variable    Coefficient       Std.Error     t-Statistic     Probability
------------------------------------------------------------------------------------
            CONSTANT     18.9104191       2.2250213       8.4989835       0.0000000
           birthrate     -0.7717761       0.1606776      -4.8032579       0.0000192
           deathrate      1.3708271       0.1293893      10.5945923       0.0000000
           migration      0.2423482       0.9605422       0.2523035       0.8020086
------------------------------------------------------------------------------------
```

Figure 8 Regression Output for Direct Factors of Population Aging

The results of the diagnostic tests are shown in the figure below.

```
REGRESSION DIAGNOSTICS
MULTICOLLINEARITY CONDITION NUMBER          44.973

TEST ON NORMALITY OF ERRORS
TEST                             DF        VALUE            PROB
Jarque-Bera                       2         1.340            0.5117

DIAGNOSTICS FOR HETEROSKEDASTICITY
RANDOM COEFFICIENTS
TEST                             DF        VALUE            PROB
Breusch-Pagan test                3         1.203            0.7522
Koenker-Bassett test              3         1.625            0.6538

SPECIFICATION ROBUST TEST
Not computed due to multicollinearity.

DIAGNOSTICS FOR SPATIAL DEPENDENCE
TEST                           MI/DF       VALUE            PROB
Moran's I (error)              0.4014       7.380            0.0000
Lagrange Multiplier (lag)         1         3.256            0.0712
Robust LM (lag)                   1         0.002            0.9663
Lagrange Multiplier (error)       1        28.755            0.0000
Robust LM (error)                 1        25.501            0.0000
Lagrange Multiplier (SARMA)       2        28.757            0.0000
```

Figure 9 Diagnostics for Regression and Spatial Dependence for Direct Factors of Population Aging

Following the diagnostics for the regression, is the diagnostics tests for spatial dependence. From the results shown in Figure 9, the Lagrange Multiplier test for error is statistically significant, that points towards a spatial error model. A spatial error occurs because error terms are correlated across observations or there are missing variables in space that are not accounted in the model (Bauhoff, 2005). To address this, one method is to model in an error term and run the regression. The results from this run in GeoDaSpace, is shown in the figure below.

```
REGRESSION
----------
SUMMARY OF OUTPUT: MAXIMUM LIKELIHOOD SPATIAL ERROR (METHOD = FULL)
-------------------------------------------------------------------
Data set             :finalmodel_var3.dbf
Weights matrix       :File: SRA_DB_JPN_knn7.gwt
Dependent Variable   :  aged_ratio              Number of Observations:          47
Mean dependent var   :     28.2851              Number of Variables   :           4
S.D. dependent var   :      2.7671              Degrees of Freedom    :          43
Pseudo R-squared     :      0.8910
Sigma-square ML      :       0.263              Log likelihood        :     -38.629
S.E of regression    :       0.512              Akaike info criterion :      85.257
                                                Schwarz criterion     :      92.658

------------------------------------------------------------------------------------
            Variable     Coefficient       Std.Error     z-Statistic     Probability
------------------------------------------------------------------------------------
            CONSTANT      26.4552224       2.1024289      12.5831708       0.0000000
           birthrate      -1.4416785       0.1490896      -9.6698778       0.0000000
           deathrate       1.1824129       0.1024576      11.5405141       0.0000000
              lambda       0.8304726       0.0768383      10.8080482       0.0000000
           migration       1.0016810       0.6074707       1.6489371       0.0991605
------------------------------------------------------------------------------------
```

Figure 10 Results of Spatial Error Model with ML Method

The spatial error term, lambda, is significant, and the Akaike Info Criterion reduces, which shows an improvement in the model. To address the question of missing variables leading to spatial errors, additional exploration and inclusion of indirect factors is conducted. The following section details the methodology and outcome for that.

Exploration of Indirect Factors of Aging Population

The dependent variable is still the ratio of population [65 years old and over]. However, for selection of explanatory variables, machine learning method for feature importance is applied. After removing variables that are highly correlated with the dependent variable, Python's sklearn library is utilized to run a tree-based regression on all the variables in the combined dataset. The top features are ranked and the top 20 are plotted below. Figures 19-28 in the appendix show a similar plot for each of the topics available in the dataset to gauge the top features by each topic.

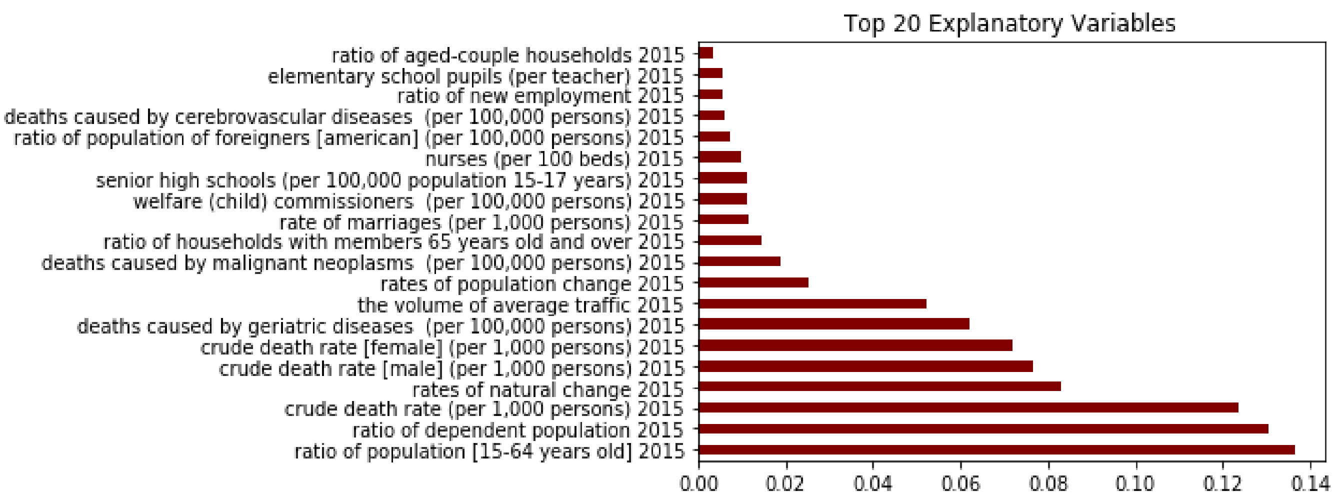


Figure 11 Feature Importance Plot using Tree Based Regression

To ensure that the highly correlated variables are not selected, a correlation matrix is analyzed to evaluate the similarity among explanatory variables and with dependent variable. The result is shown in the figure below. The highly correlated variables to the dependent variable include, deaths caused by geriatric diseases, deaths caused by malignant neoplasms, ratio of dependent population and crude death rates. While these are expected, additional themes such as education, inhabitable area, rates of marriages, migration, and welfare emerge as factors for aging concentration, indicating the presence of an underlying social effect that may be diffused spatially.

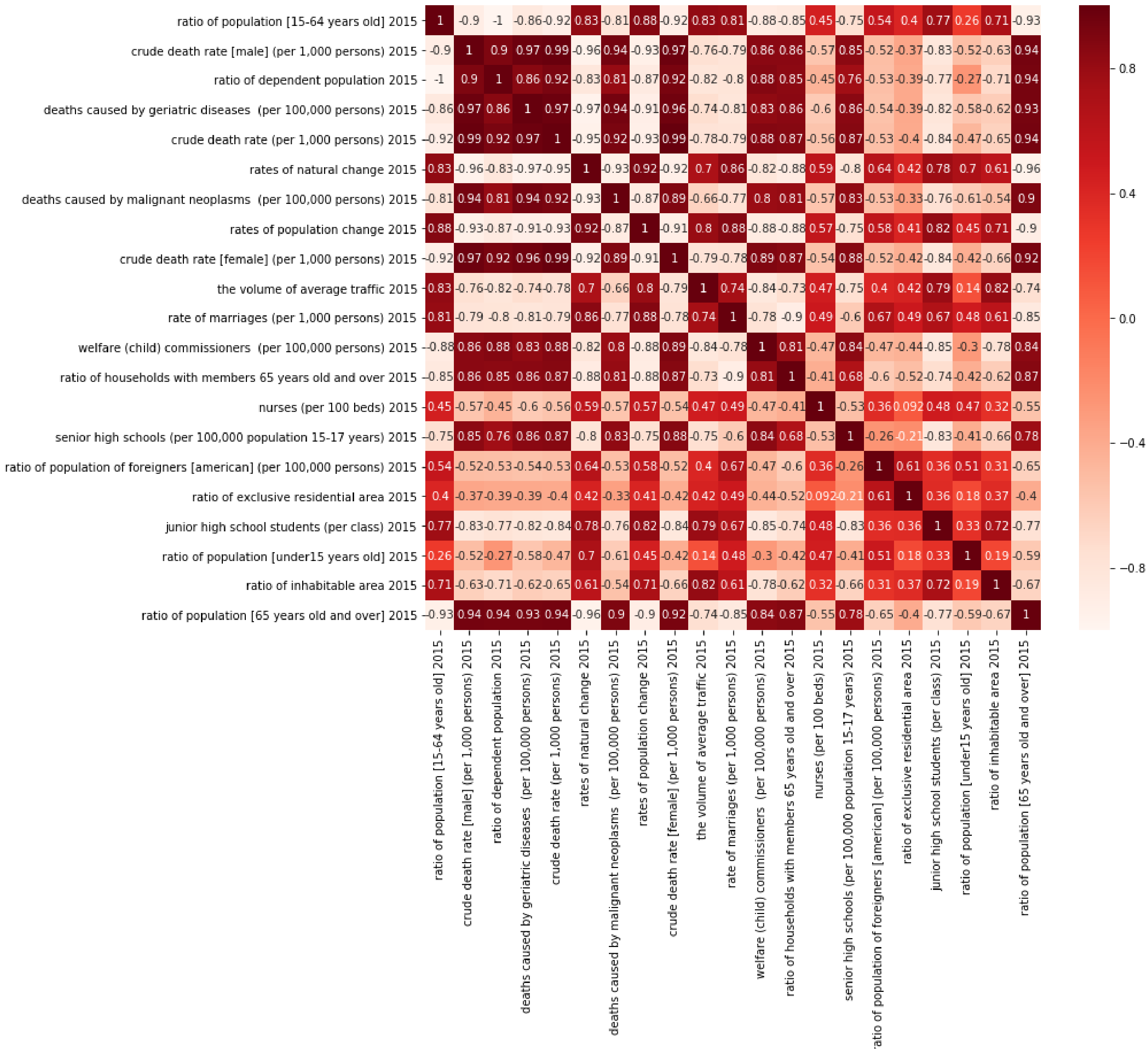


Figure 12 Correlation Plot for Top 20 Features Identified using Tree Based Regression

As discussed in the introduction, according to demographic transition theory societies see a transition from high birth rate and high mortality to low birthrate and low mortality as they prosper economically and advance technologically followed by changes in social norms. With more women in workforce, or higher cost of living and education, it is likely that people choose to have less children (Lerch, 2018). Variation on income, education, work availability and generally living in urban areas can all be factors leading to concentration in aging.

Based on the initial exploration above and theoretical understanding, the proposed model is summarized in the table 1. The dependent variable is the ratio of population [65 years old and over]. To characterize the urbanization of an area, an important variable is 'ratio of dids (densely inhabited districts) population 2015', that is the ratio of high population density districts over all districts in each prefecture. Additionally, we characterize prefecture level differences in income, education expenditure and industrialization (or lack of) characterized by ratio of people that get jobs outside of the prefecture and ratio of people that work in manufacturing. Finally, the lower fertility rate is a demographic and social characterization of more urban areas so that is included.

Table 1 GeoDaSpace Variable Name and Its Full Description

Variable	***Description***
aged_ratio	(*dependent variable*) ratio of population [65 years old and over] 2015
did_pop	ratio of dids (densely inhabited districts) population 2015
incomepp	prefectural income per person 2015
edupercap	education expenditure per capita [prefecture + municipality] 2015
jobsout	ratio of high school graduates getting jobs outside the prefecture 2015
secondind	ratio of persons employed in the secondary industry 2015
fertility	total fertility rate 2015

Following the model specification, an iterative process of spatial regression analysis is conducted in opensource software, GeoDaSpace. Like the analysis for direct factors, the method here includes an ordinary least squares regression run, diagnostic test for normality, multicollinearity and heteroskedasticity in errors and spatial diagnostic tests. The figure below is the output of the initial regression. It is seen that of all the explanatory variables, the ratio of DIDs (densely inhabited districts) population and fertility are significant, and they explain ~63% of the variation in ratio of population over the age of 65. The two significant variables are negatively related to dependent variable. From this it can be explained that a one-unit increase in ratio of densely inhabited districts leads to a 0.11 decrease in the ratio of population over the age of 65. And one-unit increase in total fertility rate, decreases the ratio of population over the age of 65 by 9.41.

```
REGRESSION
----------
SUMMARY OF OUTPUT: ORDINARY LEAST SQUARES
-----------------------------------------
Data set            :finalmodel_var3.dbf
Weights matrix      :File: SRA_DB_JPN_knn7.gwt
Dependent Variable  :  aged_ratio                Number of Observations:          47
Mean dependent var  :    28.2851                 Number of Variables   :           7
S.D. dependent var  :     2.7671                 Degrees of Freedom    :          40
R-squared           :     0.6366
Adjusted R-squared  :     0.5821
Sum squared residual:    128.001                 F-statistic           :     11.6780
Sigma-square        :      3.200                 Prob(F-statistic)     :   1.595e-07
S.E. of regression  :      1.789                 Log likelihood        :     -90.234
Sigma-square ML     :      2.723                 Akaike info criterion :     194.469
S.E of regression ML:     1.6503                 Schwarz criterion     :     207.420

------------------------------------------------------------------------------------
            Variable     Coefficient       Std.Error     t-Statistic     Probability
------------------------------------------------------------------------------------
            CONSTANT      53.2958645       7.0758104       7.5321217       0.0000000
             did_pop      -0.1198258       0.0208563      -5.7452978       0.0000011
           edupercap       0.0141216       0.0142133       0.9935460       0.3264165
           fertility      -9.4154536       2.4282606      -3.8774478       0.0003839
            incomepp      -0.0012715       0.0007560      -1.6819153       0.1003776
             jobsout      -0.0181136       0.0338465      -0.5351686       0.5954962
           secondind      -0.0969034       0.0793961      -1.2205057       0.2294196
------------------------------------------------------------------------------------
```

Figure 13 Ordinary Least Square Regression Results

The results of diagnostic tests, shown below, are used to explore model validity and spatial dependence.

```
REGRESSION DIAGNOSTICS
MULTICOLLINEARITY CONDITION NUMBER          77.525

TEST ON NORMALITY OF ERRORS
TEST                              DF        VALUE           PROB
Jarque-Bera                        2         1.063           0.5876

DIAGNOSTICS FOR HETEROSKEDASTICITY
RANDOM COEFFICIENTS
TEST                              DF        VALUE           PROB
Breusch-Pagan test                 6        14.090           0.0286
Koenker-Bassett test               6        13.108           0.0414

SPECIFICATION ROBUST TEST
Not computed due to multicollinearity.

DIAGNOSTICS FOR SPATIAL DEPENDENCE
TEST                            MI/DF       VALUE           PROB
Moran's I (error)               0.1469       3.335           0.0009
Lagrange Multiplier (lag)          1         7.619           0.0058
Robust LM (lag)                    1         3.784           0.0518
Lagrange Multiplier (error)        1         3.853           0.0497
Robust LM (error)                  1         0.017           0.8950
Lagrange Multiplier (SARMA)        2         7.636           0.0220
```

Figure 14 Diagnostic Test Results

Typically, a value under 30 is a good rule of thumb for acceptable multicollinearity condition number. However, here is it is much higher. Considering that most coefficients were insignificant, and this number is high, it is determined that the correlation among variables and dependent values should be re-evaluated to improve the model and to remove redundancy. For the test on normality, the p-value is not significant, so this indicates normality of errors. This is a good thing as the number of observations is small so non-normal data would not be suited for maximum likelihood (ML) method and with just 47 observations, generalized method of moments (GMM) methods would also not be helpful. Next, the coefficients of heteroskedasticity tests are both less than 0.05 so we reject the null which indicates some issues with heteroskedasticity. The main issue with this is, that standard errors calculated above are no longer reliable as it is required assumption for OLS. However, it is quite possible that this issue exists because of some underlying spatial dependence that is not yet accounted for, so it is reevaluated at a later stage.

From the spatial diagnostics, the Moran's I test is clearly significant, indicating strong spatial dependence. The Lagrange Multiplier tests for lag and error are both significant (less than 0.05) so we consider the Robust LM tests where both are insignificant. However, since Robust LM lag is only slightly above the threshold, a spatial lag model is the best choice. A spatial lag occurs when neighboring observations affect one another which would "lead to inconsistent and biased estimates of OLS" (Bauhoff, 2005). To address that for this small sample with no normality issues, one can run the Spatial Lag Model with maximum likelihood (ML) method (Darmafol, 2015). The results from GeoDaSpace is shown below.

```
REGRESSION
----------
SUMMARY OF OUTPUT: MAXIMUM LIKELIHOOD SPATIAL LAG (METHOD = FULL)
-----------------------------------------------------------------
Data set            :finalmodel_var3.dbf
Weights matrix      :File: SRA_DB_JPN_knn7.gwt
Dependent Variable  :  aged_ratio               Number of Observations:          47
Mean dependent var  :     28.2851               Number of Variables   :           8
S.D. dependent var  :      2.7671               Degrees of Freedom    :          39
Pseudo R-squared    :      0.7054
Spatial Pseudo R-squared:  0.6243
Sigma-square ML     :       2.210               Log likelihood        :     -86.167
S.E of regression   :       1.486               Akaike info criterion :     188.334
                                                Schwarz criterion     :     203.135

------------------------------------------------------------------------------------
            Variable     Coefficient       Std.Error     z-Statistic     Probability
------------------------------------------------------------------------------------
            CONSTANT      39.5917545       7.5498973       5.2440123       0.0000002
        W_aged_ratio       0.5028351       0.1363708       3.6872654       0.0002267
             did_pop      -0.1165809       0.0174279      -6.6893090       0.0000000
           edupercap       0.0071375       0.0118755       0.6010243       0.5478238
           fertility     -10.7943754       2.0526004      -5.2588782       0.0000001
            incomepp      -0.0010178       0.0006288      -1.6185602       0.1055419
             jobsout      -0.0119204       0.0281422      -0.4235771       0.6718742
           secondind      -0.0312378       0.0671971      -0.4648680       0.6420260
------------------------------------------------------------------------------------
```

Figure 15 Results of Spatial Lag with ML Method in GeoDaSpace

With the new spatial variable, W_aged_ratio, added to account for spatial dependence, the model's performance improves. This variable is significant and the coefficient, 0.502, is the ρ-value, a measure of spatial dependence inherent in sample data, measuring the average influence on observations by their neighboring observations. To calculate the total effect of each variable, the coefficients are multiplied by a spatial multiplier, calculated using the formula, $1/(1-\rho)$. With the ρ-value of 0.502, the multiplier is about 2. Therefore, the total effect of 1-unit increase in ratio of densely inhabited areas is 0.22 decrease in ratio of population over 65 and the total effect of such a change in fertility, is decrease in 20 unit decrease in population over 65. Of the significant variables, the coefficient for ratio of densely inhabited district is still the same, however, the negative correlation of total fertility rate increased by a little. Finally, a quick run for diagnostic tests with the spatial variable, shows that the heteroskedasticity on errors is no longer an issue. Those results are included in appendix, figure 29.

Discussion

From the initial mapping for hypothesis generation to exploratory spatial data analysis, it evident that concentration of aging population is distributed unevenly. The exploratory spatial analysis shows the clustering in the ratio of population over the age of 65 along the urban-rural divide. This is confirmatory to propositions of demographic transition theory that recognizes that technology, modernization, and urbanization play a critical role in affecting the direct factors of population transformation – birth rate, death rate and migration. To explore the impact of the direct and indirect factors leading to spatial heterogeneity of aging population, spatial econometric models are applied to each one of them.

The results of the spatial regression on direct factors, shows that birth rate, death rate and migration can explain ~92% of variation in the ratio of population over the age of 65 but have spatial error, indicating missing variables or correlation in the error terms. On the other hand, simply considering indirect factors of urbanization such as, education, fertility, densely inhabited districts, and industry, explains only ~63% of the variation in ratio of population over the age of 65 and in the model specified here, only the variables, fertility, densely inhabited districts, are significant. However, the spatial dependency of these factors is clear. The model with indirect factors is diagnosed to have missing spatial lag term. When spatial variable is added, the coefficient relation and the number of significant variables remains the same, but model's performance improves based on Akaike Info Criterion. The coefficient of spatial term is used to calculate the spatial multiplier of value 2, that is, this analysis shows that the actual effect of significant variables, densely inhabited districts (DID) and fertility is double that of it its apparent effect.

The spatial processes beneath the observations in each of the prefectures are not independent and there is an interrelation between neighboring regions which needs to be captured for the true impact. The existence of spatial effect due to the variables of densely inhabited districts (DID) and fertility is reasonable. DID captures the essence of urban areas: due to the economic and social opportunities in urban areas, growing urban regions tend to expand to nearby cities and younger people would be inclined to move into those areas, affecting the ratio of aging population in a given prefecture as well as the surrounding areas. Similarly, the lifestyle choices are based on social norms or work needs of rural and urban area and ideas such as having more or less children would diffuse in surrounding areas, also affecting the ratio of aging population.

A notable challenge of this analysis was the limited number of prefectures. With just 47 prefectures, administrative units, the explanatory variables need to be carefully selected to ensure enough degrees of freedom and statistical significance. Additionally, an important practice in spatial analysis is to vary the spatial scale determine the robustness of the results and capture any impact of the Modifiable Areal Unit Problem (MAUP)" (Openshaw, 1984). However, in this study, there is a lack of data on the municipality and town level and aggregating the data on to conduct a metropolitan area level analysis, would limit the data even further to just 11 units. Nonetheless, the model stayed consistent when the number of neighbors were varied in k-NN weight matrix and given the scale of urbanization process, the prefecture level still seems appropriate for this study.

In this study, the selection process for the limited variables was both analytical and experimental, with much room for improvement. The application of machine learning method to determine top 20 features is helpful in exploration of themes for variables to consider. However, this feature selection method resulted in highly correlated variables or selection that may be predictive but not explanatory. The figure 30 in the appendix shows the results of a regression model using 6 of the of top features. The resulting model had mostly insignificant coefficients and no spatial dependence. To find a meaningful relation, the final model for indirect factors relied heavily on the themes from feature selection as well as theory. In a final attempt to improve the results, a regression is run on the significant direct factors of birthrate and death rate and significant indirect effects, densely inhabited districts (DID) and fertility, which resulted in high R-squared value of ~93% but once again flagged spatial error. The output for this is shown in figure 31 of the appendix. While this result is not definitive, it shows the potential to understand the nuanced spatial characteristics of processes that explain the local clusters of aging population.

Finally, multicollinearity is consistently an issue with most model iterations, including, the final models. The correlation map below shows the tendency of correlation for explanatory variables and dependent variables for the indirect factors model.

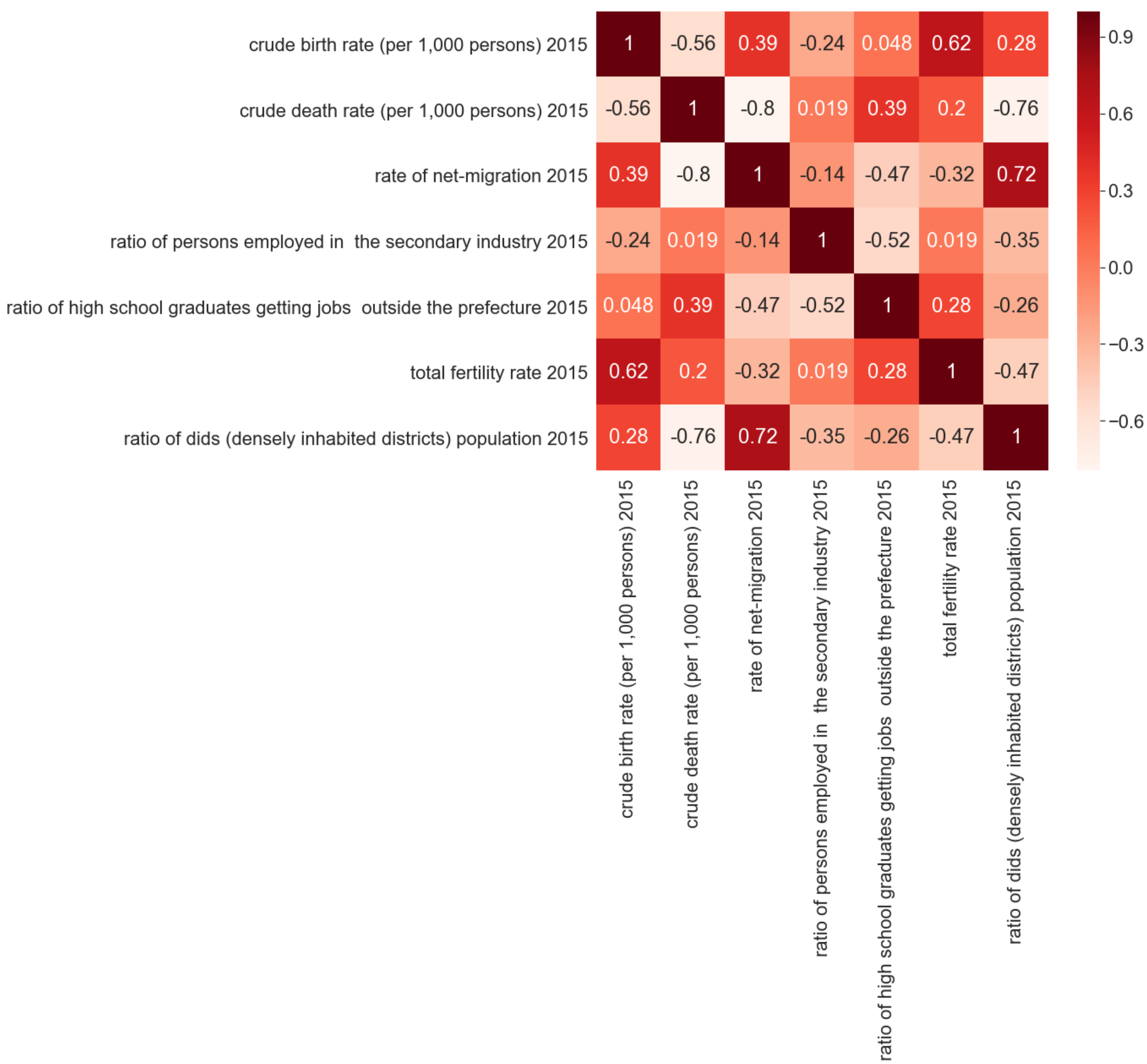


Figure 16 Correlation Map of Explanatory variables and Dependent Variables

From the figure above, it is apparent that there is a strong correlation between the ratio of aging population over the age of 65 and crude death rate and rate of migration and ratio of DIDs population. Guided by the correlation matrix and additional variance inflation factor (VIF) analysis, the model is modified to minimize the issue on multicollinearity. While it is not always successful, a restricted model with just ratio of DIDs (densely inhabited districts) population and total fertility rate, shown in figure 36, did see an improvement in this aspect. This suggests that a strong model with just 2 to 3 relevant variables would be more suitable for the spatial analysis of this small nation. Figures 32 to 37 in the appendix show different iterations of the model with variations with restricted variable selection.

Conclusion and Future Work

This study explores the spatial heterogeneity in the ratio of population over the age of 65 in Japan and the direct and indirect factors that contribute to the patterns in concentration of aging population. An initial exploration shows that urban, industrialized areas have a lower concentration of aging population compared to less urban, industrialized area. The paper first utilizes population dynamics theory, ordinary least square regression as well as spatial regression to understand the effect of direct factors of population aging. The classic factors of population aging, birthrate, death rate and migration, have a high predictive ability, but it is missing critical understanding of social and urban factors. The paper then uses machine learning methods, exploratory spatial analysis, and spatial econometrics to understand the effect of the indirect factors on population aging. From the results of the specified model, it is seen that the indirect factors, the ratio of DIDs (densely inhabited districts) population and total fertility rate, results in a lower performing model than direct factors. However, that model with spatial lag results in a spatial multiplier of 2, that indicates that the direct effects of the selected urban features is double the apparent effect.

Understanding that the direct factors of population dynamics are driven by indirect factors of urbanization and modernization and being able to spatially analyze and quantify it is useful in developing strong regional and national policies. To mitigate the impact of rapid aging in the regions, the solution may be better infrastructure connecting nearby areas with high commercial centers, policy changes that improves work life balance or industrialization. The solution should use the theoretical and analytical framework of such a spatial analysis combined with knowledge of the culture and strengths of the area in consideration.

The model for indirect effects indicates the potential for exploring the spatial effects in population aging, but this study can be improved by addressing the issues of multicollinearity and statistical significance of variables by employing more discerning methods for feature selection. More sophisticated spatial econometric methods, such as using instrument variables, can also be utilized to improve the significance of the model results. Finally, a deeper qualitative understanding of the country's culture and institutional structures would certainly guide the model selection better. To aid the reproduction and continuation of this study, all data, code and output is available here.

References

Anselin, L., Syabri, I., & Kho, Y. (2009). GeoDa: An Introduction to Spatial Data Analysis. *Handbook of Applied Spatial Analysis,* 73-89. doi:10.1007/978-3-642-03647-7_5

Bauhoff, S. (2005, November 15). Spatial Error. Retrieved August 07, 2020, from https://blogs.iq.harvard.edu/spatial_error_1

Bauhoff, S. (2005, October 11). Spatial Lag. Retrieved August 07, 2020, from https://blogs.iq.harvard.edu/spatial_lag_1

Blanford, J., Kessler, F., Griffin, A., & O'Sullivan, D. (2018). Project 4: Calculating Global Moran's I and the Moran Scatterplot. Retrieved June 08, 2020, from https://www.e-education.psu.edu/geog586/node/672

Darmofal, D. (2015). *Spatial analysis for the social sciences*. New York, NY: Cambridge University Press.

Davies, A., & James, A. (2016). *Geographies of Ageing Social Processes and the Spatial Unevenness of Population Ageing*. London: Taylor and Francis.

GADM Japan. (2018). Retrieved November 30, 2019, from https://gadm.org/maps/JPN.html

Galor, O. (2011). The Demographic Transition: Causes and Consequences. doi:10.3386/w17057

Galor, O., & Weil, D. N. (2000). Population, Technology, and Growth: From Malthusian Stagnation to the Demographic Transition and Beyond. *American Economic Review, 90*(4), 806-828. doi:10.1257/aer.90.4.806

Grafeneder-Weissteiner, T., Prettner, K., & Südekum, J. (2020). Three Pillars of Urbanization: Migration, Aging, and Growth. *De Economist, 168*(2), 259-278. doi:10.1007/s10645-020-09356-z

Harper, S., & Laws, G. (1995). Rethinking the geography of ageing. *Progress in Human Geography, 19*(2), 199-221. doi:10.1177/030913259501900203

Kawabe, H. (1980). The population trend and the population problems in recent Japan. *GeoJournal, 4*(3). doi:10.1007/bf00218575

Kim Parker, J. (2020, May 30). Similarities and differences between urban, suburban and rural communities in America. Retrieved July 31, 2020, from https://www.pewsocialtrends.org/2018/05/22/what-unites-and-divides-urban-suburban-and-rural-communities/

Kirk, D. (1996). Demographic Transition Theory. *Population Studies, 50*(3), 361-387. doi:10.1080/0032472031000149536

Lagerlof, N. (2000). From Malthus to Modern Growth: The Three Regimes Revisited.

Lerch, M. (2018). Fertility Decline in Urban and Rural Areas of Developing Countries. *Population and Development Review, 45*(2), 301-320. doi:10.1111/padr.12220

Matthews, S. A., & Parker, D. M. (2013). Progress in Spatial Demography. *Demographic Research, 28*, 271-312. doi:10.4054/demres.2013.28.10

Measuring Age Friendliness of Cities: A Guide to Using Core Indicators (pp. 1-128, Publication). (n.d.). World Health Organization.

Michael, R. (2018, May 15). Fukuoka. Retrieved August 07, 2020, from https://www.britannica.com/place/Fukuoka-Japan

Nakamura, D. (2013). Spatial Policy for a Competitive Regional System: Economic and Social Infrastructure Elements. *Journal of Urban Management, 2*(1), 103-112. doi:10.1016/s2226-5856(18)30067-0

Nishida, T., Pick, J. B., & Sarkar, A. (2014). Japanʹs prefectural digital divide: A multivariate and spatial analysis. *Telecommunications Policy, 38*(11), 992-1010. doi:10.1016/j.telpol.2014.05.004

Notestein, F. W. (1954). Introduction. In *Some demographic aspects of aging* (pp. 38-45). Philadelphia, PA: American philosophical Society.

OECD Territorial Reviews: Japan 2016 (Publication). (2016). OECD Territorial Reviews. doi:doi:10.1787/9789264250543-en

Olshansky, S. J. (1997). *The demography of aging*. Silver Spring, MD, MD: Population Association of America.

Pletcher, K. (2013, June 11). Yamanashi. Retrieved August 07, 2020, from https://www.britannica.com/place/Yamanashi

Pletcher, K. (2013, May 17). Hiroshima. Retrieved August 07, 2020, from https://www.britannica.com/place/Hiroshima-prefecture-Japan

Preston, S. H., Himes, C., & Eggers, M. (1989). Demographic Conditions Responsible for Population Aging. *Demography, 26*(4), 691. doi:10.2307/2061266

Shimada, T. (2004). Spatial Diffusion of Residential Burglaries in Tokyo: Using Exploratory Spatial Data Analysis. *Behaviormetrika, 31*(2), 169-181. doi:10.2333/bhmk.31.169

Shiode, N., Morita, M., Shiode, S., & Okunuki, K. (2014). Urban and rural geographies of aging: A local spatial correlation analysis of aging population measures. *Urban Geography, 35*(4), 608-628. doi:10.1080/02723638.2014.905256

Szołtysek, M., Ogórek, B., Poniat, R., & Gruber, S. (2019). Making a Place for Space: A Demographic Spatial Perspective on Living Arrangements Among the Elderly in Historical Europe. *European Journal of Population, 36*(1), 85-117. doi:10.1007/s10680-019-09520-5

Szreter, S. (1993). The Idea of Demographic Transition and the Study of Fertility Change: A Critical Intellectual History. *Population and Development Review, 19*(4), 659. doi:10.2307/2938410

West, J. (2014, July 09). JAPAN'S DEMOGRAPHIC DRAMA. Retrieved November 06, 2019, from https://asiancenturyinstitute.com/society/443-japan-s-demographic-drama

Wu, Y., Song, Y., & Yu, T. (2019). Spatial Differences in China's Population Aging and Influencing Factors: The Perspectives of Spatial Dependence and Spatial Heterogeneity. *Sustainability, 11*(21), 5959. doi:10.3390/su11215959

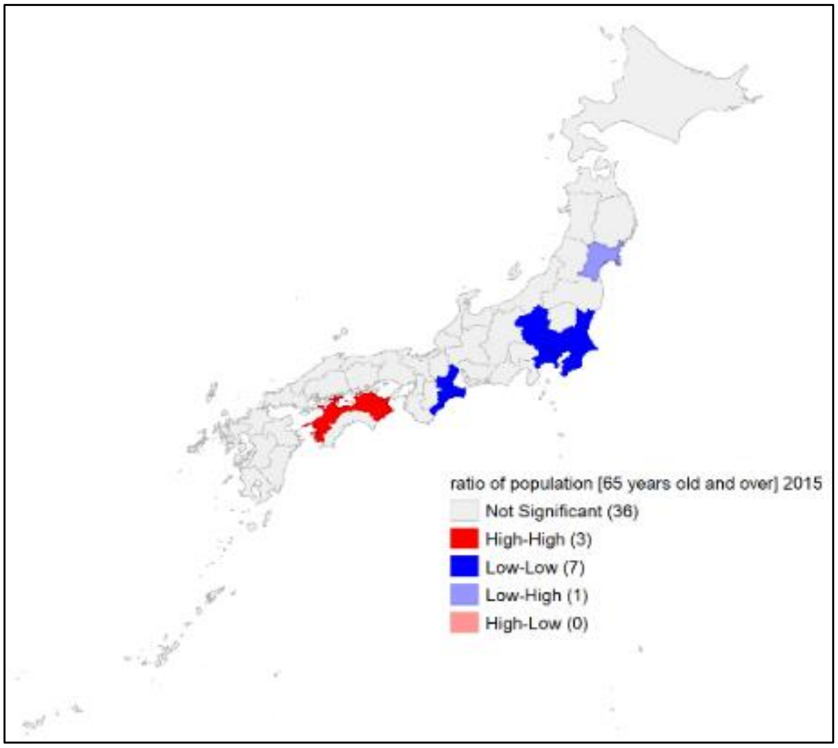
Xie, B., Zhou, J., & Luo, X. (2015). Mapping spatial variation of population aging in China's mega cities. *Journal of Maps, 12*(1), 181-192. doi:10.1080/17445647.2014.1000984

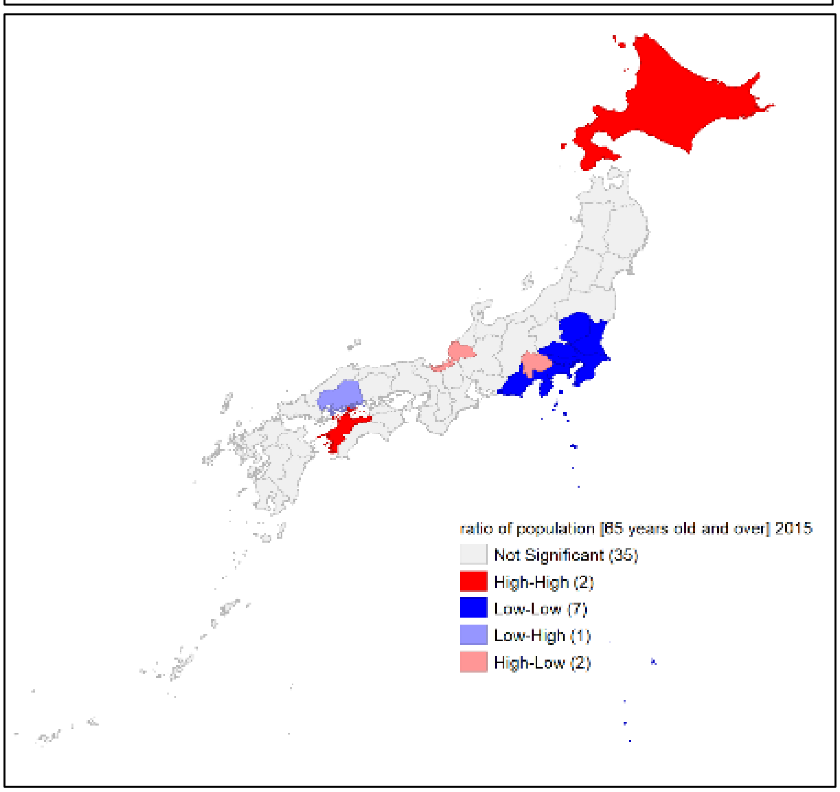
Xu, X., Zhao, Y., Zhang, X., & Xia, S. (2018). Identifying the Impacts of Social, Economic, and Environmental Factors on Population Aging in the Yangtze River Delta Using the Geographical Detector Technique. *Sustainability, 10*(5), 1528. doi:10.3390/su10051528

Appendix

Exploring clusters through LISA

Varying the number of neighbors in k-NN varies in local clusters in aging population.





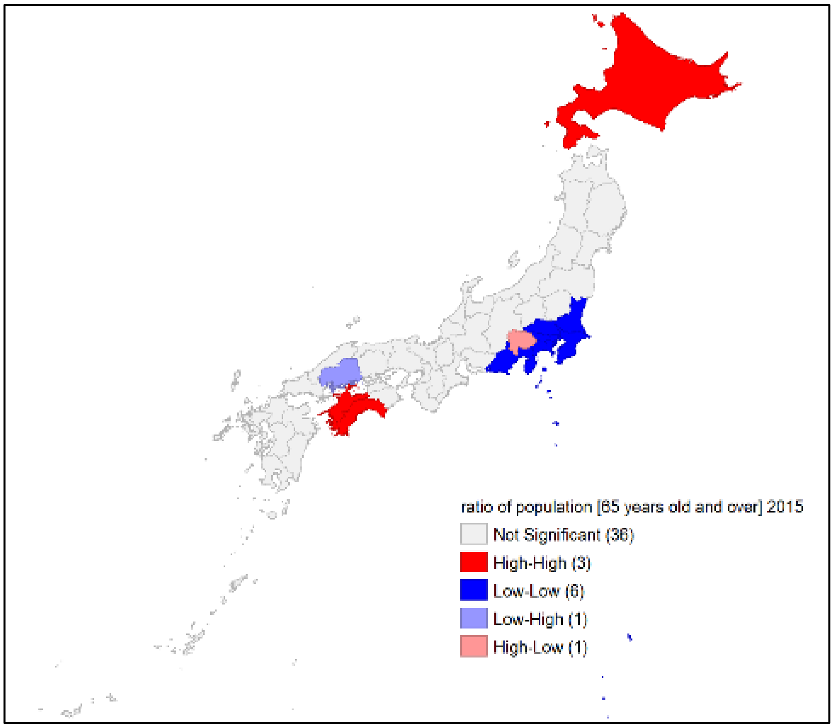
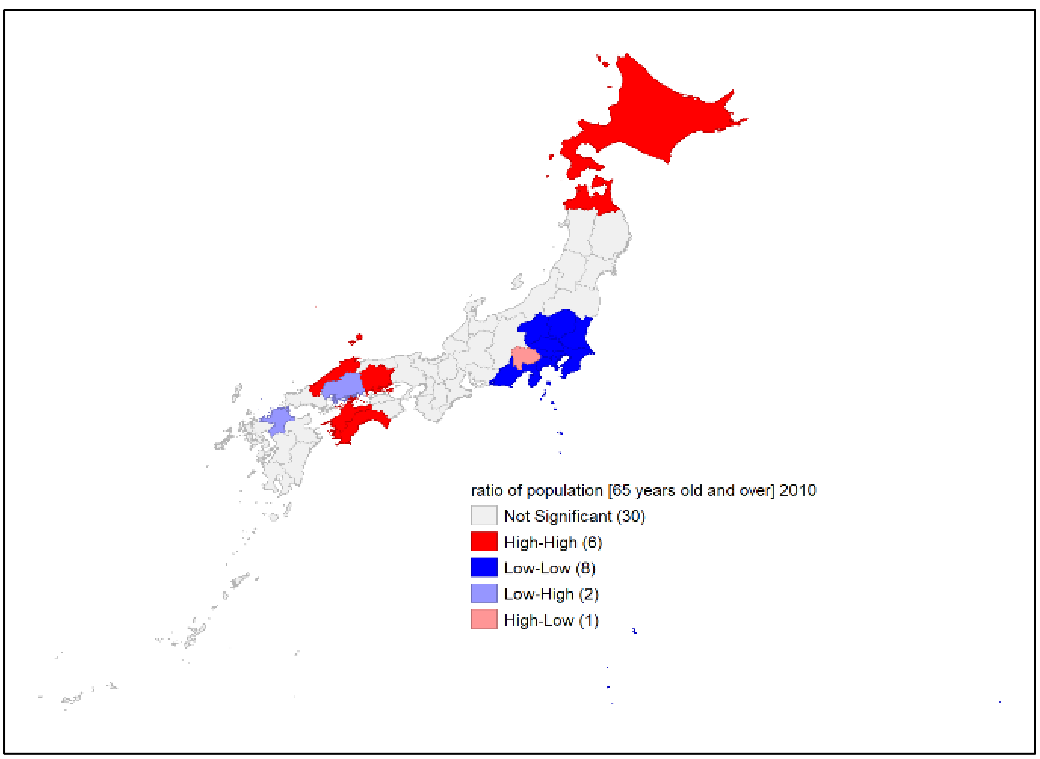


Figure 17 Local Indicators of Spatial Autocorrelation KNN = 4,5, 6, $p < 0.05$ for 2015

LISA on ratio of aging population in 2010



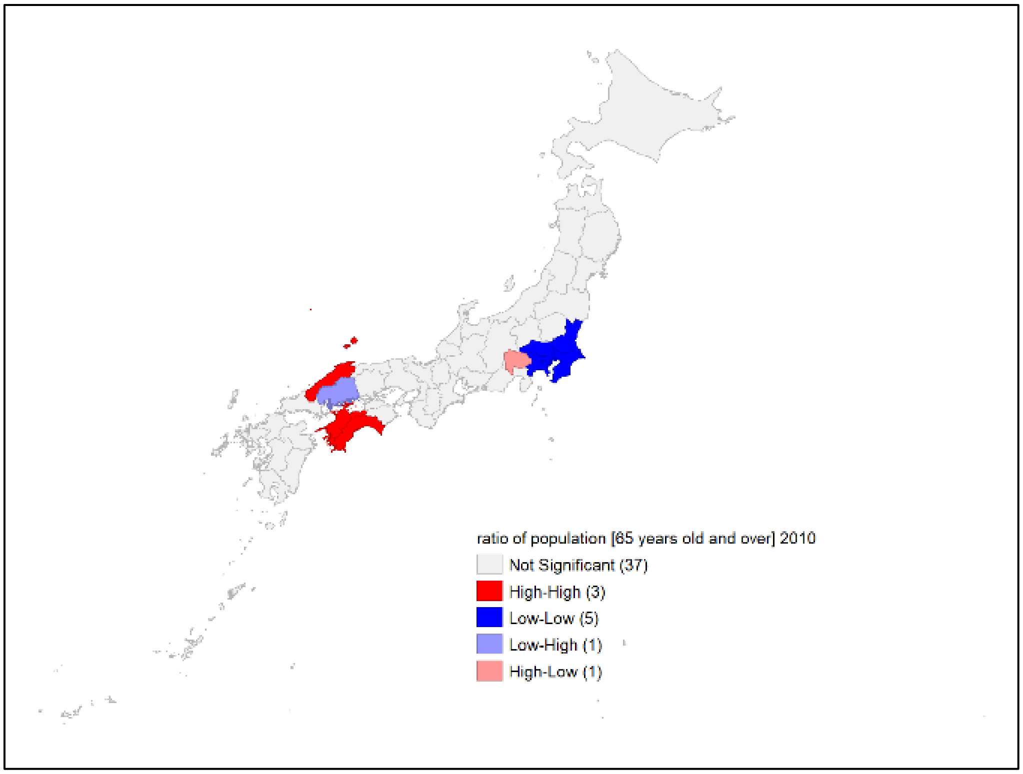


Figure 18 Local Indicators of Spatial Autocorrelation KNN = 7: p < 0.05 and P <0.01 2010

Individual Topic Top Features

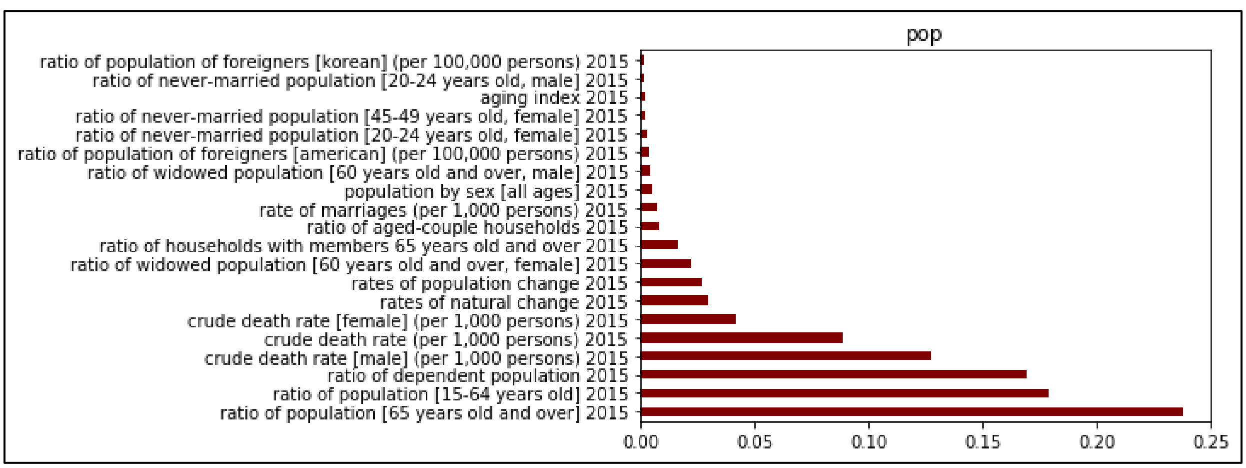


Figure 19 Feature Importance Plot using Tree Based Regression for Population Variables

Figure 20 Feature Importance Plot using Tree Based Regression for Administrative Base Variables

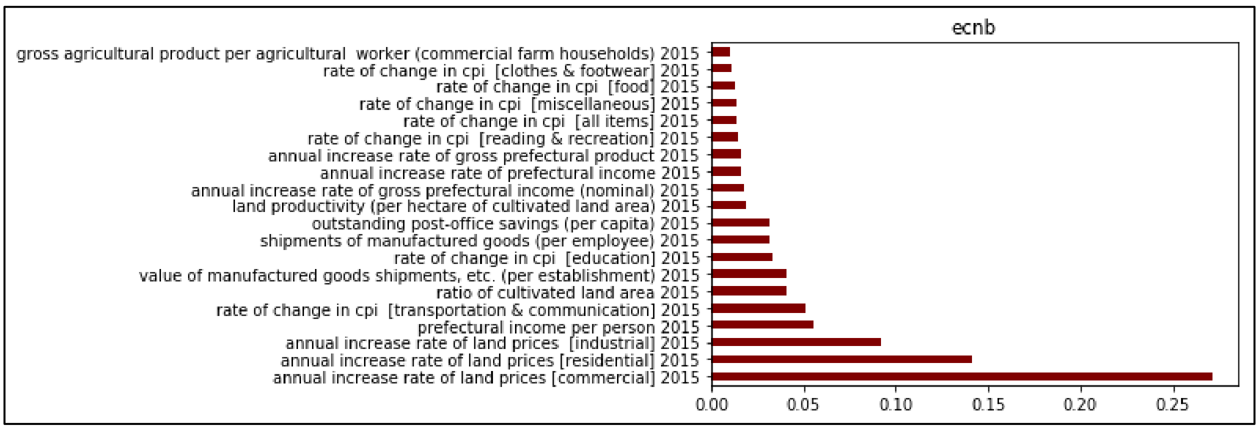


Figure 21 Feature Importance Plot using Tree Based Regression for Economic Base Variables

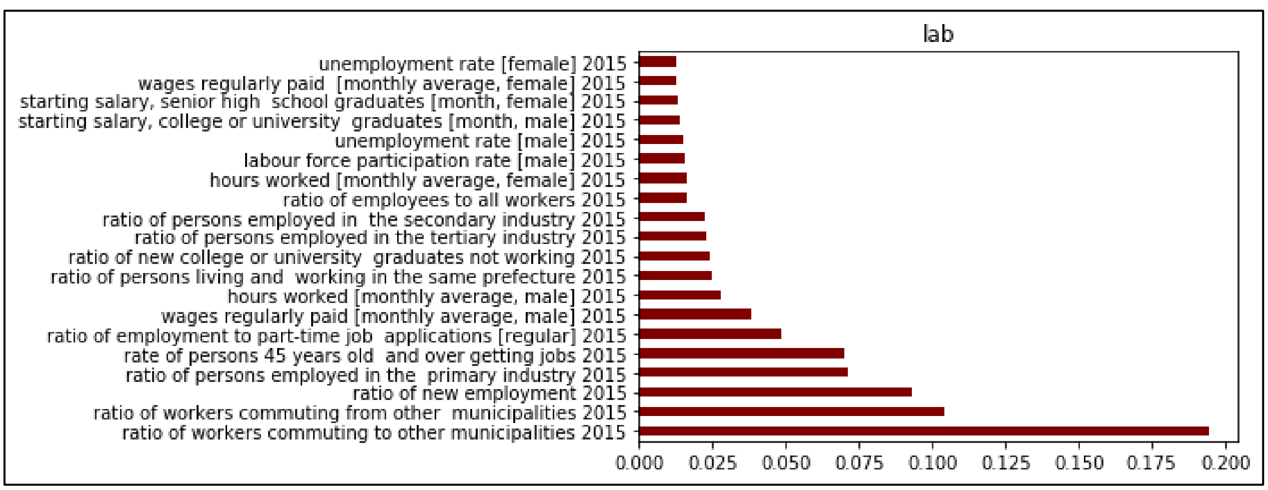


Figure 22 Feature Importance Plot using Tree Based Regression for Labor Variables

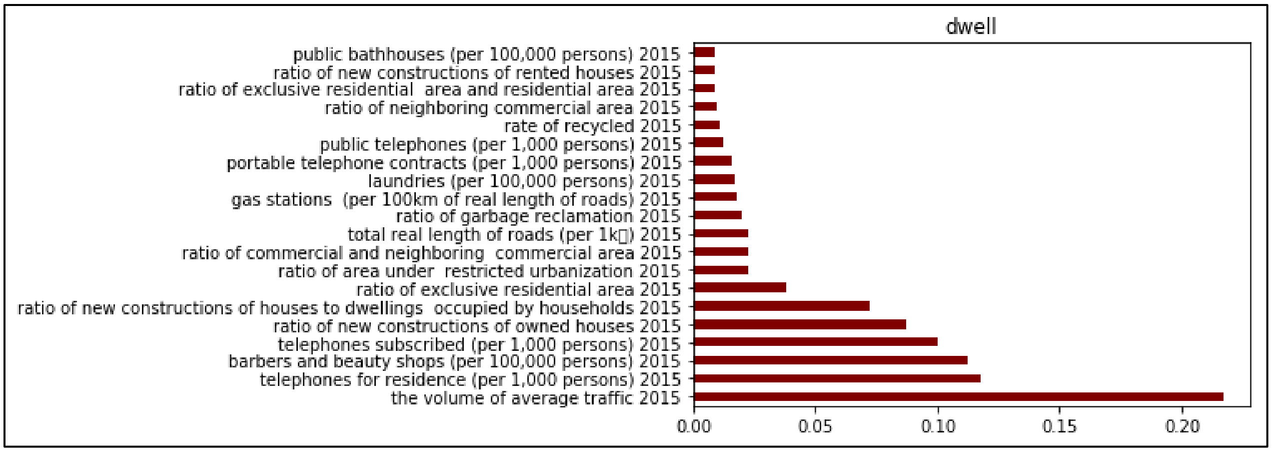


Figure 23 Feature Importance Plot using Tree Based Regression for Variables on Dwelling

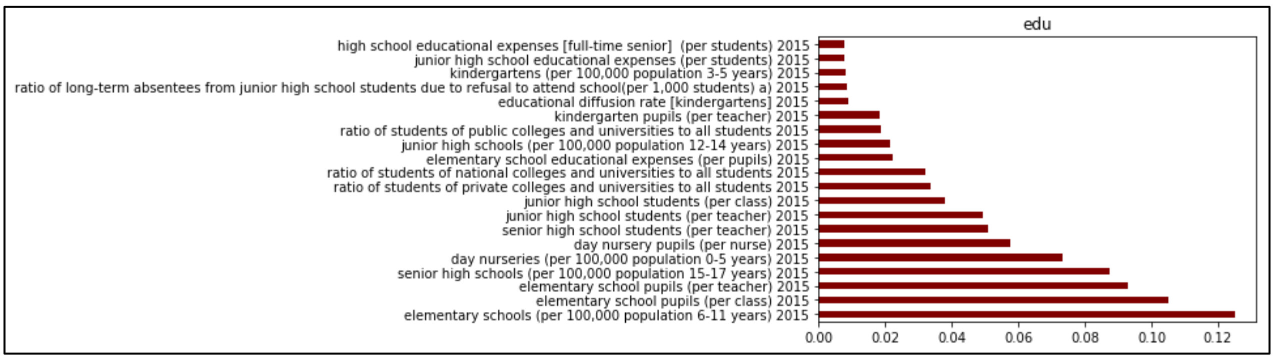


Figure 24 Feature Importance Plot using Tree Based Regression for Education Variables

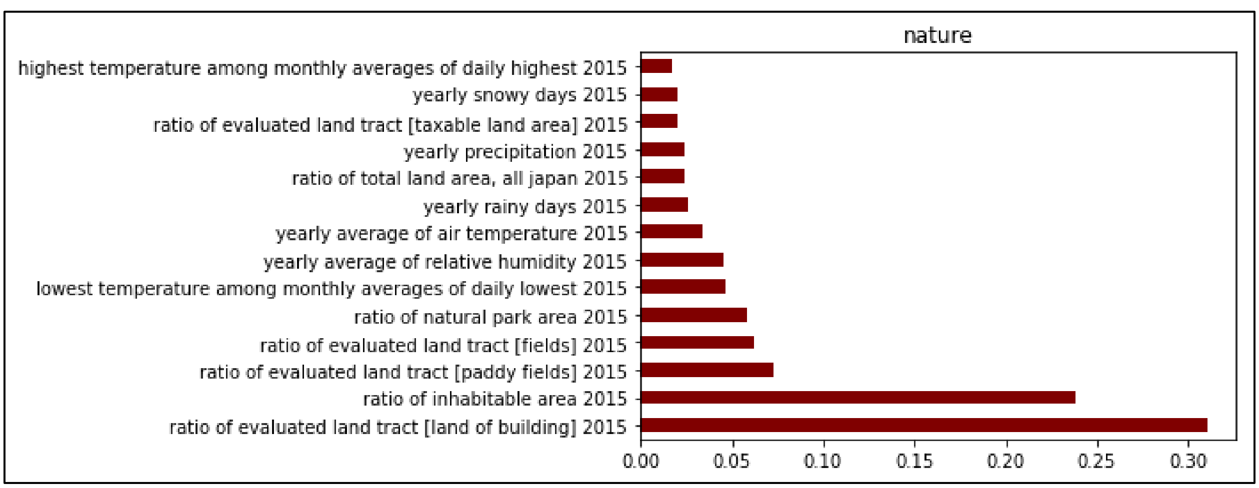


Figure 25 Feature Importance Plot using Tree Based Regression for Natural Environment Variables

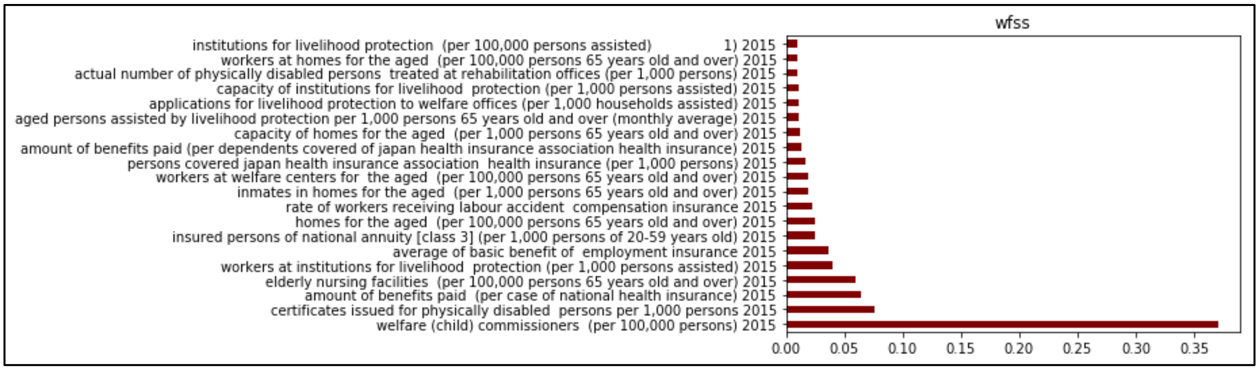


Figure 26 Feature Importance Plot using Tree Based Regression for Welfare and Social Security Variables

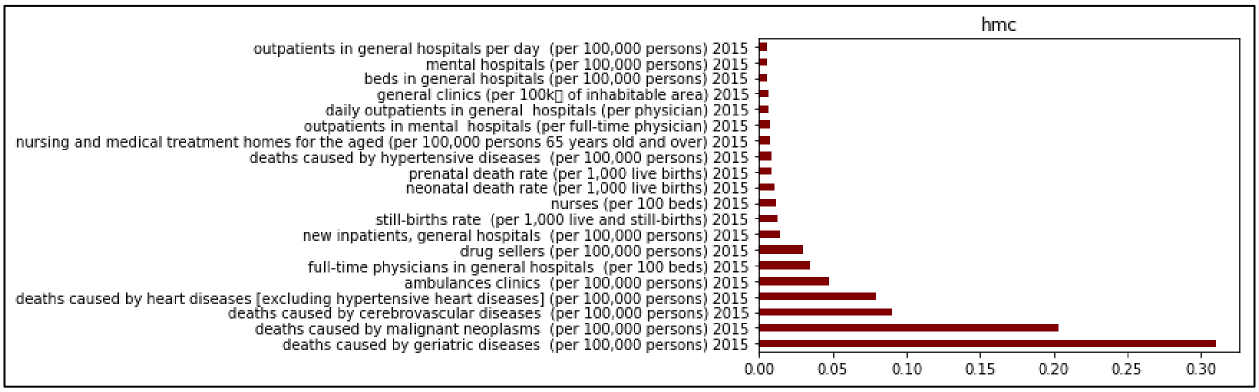


Figure 27 Feature Importance Plot using Tree Based Regression for Health and Medical Care Variables

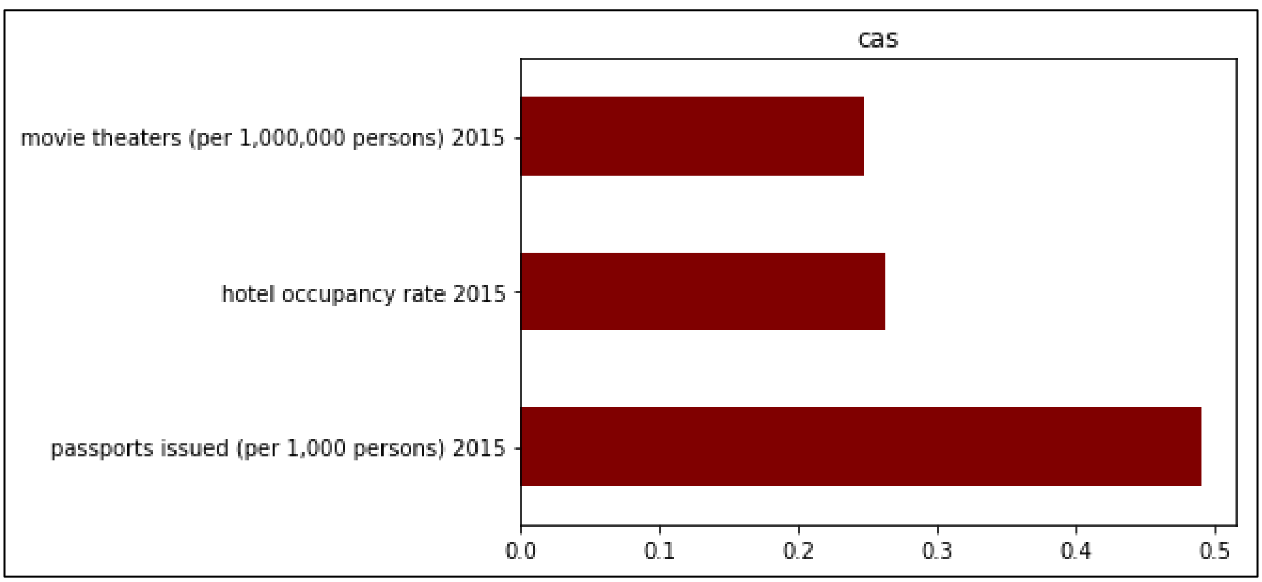


Figure 28 Feature Importance Plot using Tree Based Regression for Cultural Variables

Tests for Spatial Dependence

Table 2 Description and Passing Criteria for Tests Conducted

Name of the Test	***Description***	***Reject Null if***
Multicollinearity Condition Number	Is the correlation among variables acceptable?	Less than 30
Jarque-Bera Test	Normality on errors	$p < 0.40$
Bruesh-Pagan Test	Heteroskedasticity	$p < 0.05$
Koenker-Bassett Test	Heteroskedasticity	$p < 0.05$
Spatial Diagnostic Tests		
Moran's I (error)	To test for presence of spatial terms	$p < 0.05$
Lagrange Multiplier (lag)	To test for spatial lag	$p < 0.05$
Robust Lagrange Multiplier (lag)	To test for spatial lag if both LM lag and error are insignificant	$p < 0.05$
Lagrange Multiplier (error)	To test for spatial lag	$p < 0.05$
Robust Lagrange Multiplier (error)	To test for spatial lag if both LM lag and error are insignificant	$p < 0.05$

Rerun of Spatial Regression with Spatial Variable

```
REGRESSION
----------
SUMMARY OF OUTPUT: ORDINARY LEAST SQUARES
-----------------------------------------
Data set            :finalmodel_var2.dbf
Weights matrix      :        None
Dependent Variable  :  aged_ratio                Number of Observations:          47
Mean dependent var  :     28.2851                Number of Variables   :           8
S.D. dependent var  :      2.7671                Degrees of Freedom    :          39
R-squared           :      0.7089
Adjusted R-squared  :      0.6566
Sum squared residual:     102.542                F-statistic           :     13.5658
Sigma-square        :       2.629                Prob(F-statistic)     :   9.868e-09
S.E. of regression  :       1.622                Log likelihood        :     -85.023
Sigma-square ML     :       2.182                Akaike info criterion :     186.046
S.E of regression ML:      1.4771                Schwarz criterion     :     200.847

White Standard Errors
-----------------------------------------------------------------------------------
            Variable     Coefficient       Std.Error     t-Statistic    Probability
-----------------------------------------------------------------------------------
            CONSTANT      35.5659617       6.3792531       5.5752549      0.0000020
         W_aged_rati       0.6505507       0.2061062       3.1563865      0.0030764
             did_pop      -0.1156277       0.0157043      -7.3627831      0.0000000
           edupercap       0.0050858       0.0112060       0.4538473      0.6524543
           fertility     -11.1994549       3.1973727      -3.5027055      0.0011716
            incomepp      -0.0009432       0.0003950      -2.3877827      0.0218917
             jobsout      -0.0101011       0.0277072      -0.3645648      0.7174057
           secondind      -0.0119475       0.0761643      -0.1568654      0.8761602
-----------------------------------------------------------------------------------

REGRESSION DIAGNOSTICS
MULTICOLLINEARITY CONDITION NUMBER          113.909

TEST ON NORMALITY OF ERRORS
TEST                             DF        VALUE           PROB
Jarque-Bera                       2           0.840           0.6570

DIAGNOSTICS FOR HETEROSKEDASTICITY
RANDOM COEFFICIENTS
TEST                             DF        VALUE           PROB
Breusch-Pagan test                7           3.950           0.7856
Koenker-Bassett test              7           3.737           0.8095

DIAGNOSTICS FOR SPATIAL DEPENDENCE
TEST                           MI/DF       VALUE           PROB
Lagrange Multiplier (lag)         1           0.000           1.0000
Robust LM (lag)                   1           0.006           0.9409
Lagrange Multiplier (error)       1           0.002           0.9643
Robust LM (error)                 1           0.008           0.9310
Lagrange Multiplier (SARMA)       2           0.008           0.9963
```

Figure 29 Rerun OLS with a spatial variable

Alternative Iterations of Model Development

Model using top features from Feature Selection Machine Learning:

```
SUMMARY OF OUTPUT: ORDINARY LEAST SQUARES ESTIMATION
Data set            :  gadm36_JPN_1
Dependent Variable  :  ratio of population [65 years old and over] 2015
Number of Observations:   47
Mean dependent var  :     28.2851  Number of Variables    :    7
S.D. dependent var  :     2.73752  Degrees of Freedom     :   40

R-squared           :    0.744304  F-statistic            :      19.406
Adjusted R-squared  :    0.705950  Prob(F-statistic)      :1.88705e-010
Sum squared residual:     90.0611  Log likelihood         :    -81.9731
Sigma-square        :     2.25153  Akaike info criterion :     177.946
S.E. of regression  :     1.50051  Schwarz criterion      :     190.897
Sigma-square ML     :     1.91619
S.E of regression ML:     1.38427

-----------------------------------------------------------------------------
       Variable      Coefficient      Std.Error    t-Statistic   Probability
-----------------------------------------------------------------------------
          CONSTANT       17.0511        5.88985        2.89499     0.00612
the vol..ffic 2015  -0.000239027    0.000124258       -1.92365     0.06154
ratio o..ties 2015     0.0760085      0.0484818        1.56777     0.12481
full-ti..eds) 2015    -0.0866909       0.169679      -0.510911     0.61222
telepho..ons) 2015     0.0573957       0.015239        3.76638     0.00053
ratio o..ment 2015      0.452346       0.248023        1.82381     0.07566
ratio o..area 2015    -0.0638933      0.0320589         -1.993     0.05311
-----------------------------------------------------------------------------

REGRESSION DIAGNOSTICS
MULTICOLLINEARITY CONDITION NUMBER   78.218832
TEST ON NORMALITY OF ERRORS
TEST                  DF           VALUE             PROB
Jarque-Bera            2            0.6262           0.73118

DIAGNOSTICS FOR HETEROSKEDASTICITY
RANDOM COEFFICIENTS
TEST                  DF           VALUE             PROB
Breusch-Pagan test     6            13.5344          0.03529
Koenker-Bassett test   6            12.2289          0.05705
SPECIFICATION ROBUST TEST
TEST                  DF           VALUE             PROB
White                 27            39.2834          0.05970

DIAGNOSTICS FOR SPATIAL DEPENDENCE
FOR WEIGHT MATRIX : SRA_JPN_wt_knn4
   (row-standardized weights)
TEST                           MI/DF        VALUE           PROB
Moran's I (error)              0.0853        1.4471          0.14786
Lagrange Multiplier (lag)        1           0.0821          0.77451
Robust LM (lag)                  1           1.0379          0.30832
Lagrange Multiplier (error)      1           0.7868          0.37508
Robust LM (error)                1           1.7426          0.18681
Lagrange Multiplier (SARMA)      2           1.8246          0.40159
```

Figure 30 Model from selected features from feature selection

Explanatory Variables:

- 'the volume of average traffic'
- 'ratio of workers commuting to other municipalities'
- 'full-time physicians in general hospitals (per 100 beds)'
- 'telephones subscribed (per 1,000 persons)',
- 'ratio of new employment'
- 'ratio of exclusive residential area'

Model with significant direct and indirect factors of population aging

```
REGRESSION
----------
SUMMARY OF OUTPUT: ORDINARY LEAST SQUARES
-----------------------------------------
Data set            :finalmodel_var3.dbf
Weights matrix      :File: SRA_DB_JPN_knn7.gwt
Dependent Variable  :  aged_ratio              Number of Observations:      47
Mean dependent var  :     28.2851              Number of Variables   :       5
S.D. dependent var  :      2.7671              Degrees of Freedom    :      42
R-squared           :      0.9372
Adjusted R-squared  :      0.9312
Sum squared residual:      22.120              F-statistic           :   156.6924
Sigma-square        :       0.527              Prob(F-statistic)     :  1.183e-24
S.E. of regression  :       0.726              Log likelihood        :    -48.979
Sigma-square ML     :       0.471              Akaike info criterion :    107.958
S.E of regression ML:      0.6860              Schwarz criterion     :    117.209

-----------------------------------------------------------------------------------
            Variable     Coefficient       Std.Error     t-Statistic     Probability
-----------------------------------------------------------------------------------
            CONSTANT      16.3308227       2.6062300       6.2660712       0.0000002
           birthrate      -1.7505941       0.3936286      -4.4473246       0.0000626
           deathrate       1.1972514       0.1468254       8.1542515       0.0000000
             did_pop       0.0284664       0.0116248       2.4487772       0.0185900
           fertility       6.9368108       2.4832010       2.7934955       0.0078207
-----------------------------------------------------------------------------------
```

```
REGRESSION DIAGNOSTICS
MULTICOLLINEARITY CONDITION NUMBER          102.952

TEST ON NORMALITY OF ERRORS
TEST                                 DF        VALUE             PROB
Jarque-Bera                           2          1.400             0.4965

DIAGNOSTICS FOR HETEROSKEDASTICITY
RANDOM COEFFICIENTS
TEST                                 DF        VALUE             PROB
Breusch-Pagan test                    4          0.995             0.9105
Koenker-Bassett test                  4          0.843             0.9325

SPECIFICATION ROBUST TEST
Not computed due to multicollinearity.

DIAGNOSTICS FOR SPATIAL DEPENDENCE
TEST                               MI/DF       VALUE             PROB
Moran's I (error)                  0.2820        5.371             0.0000
Lagrange Multiplier (lag)             1          4.810             0.0283
Robust LM (lag)                       1          0.853             0.3558
Lagrange Multiplier (error)           1         14.191             0.0002
Robust LM (error)                     1         10.233             0.0014
Lagrange Multiplier (SARMA)           2         15.043             0.0005
```

Figure 31 Model Combining Direct and Indirect Factors of Population Aging

Model with birth rate and death rate. R-squared 92%, Spatial error.

```
REGRESSION
----------
SUMMARY OF OUTPUT: ORDINARY LEAST SQUARES
-----------------------------------------
Data set            :finalmodel_var3.dbf
Weights matrix      :File: SRA_DB_JPN_knn7.gwt
Dependent Variable  :  aged_ratio              Number of Observations:       47
Mean dependent var  :     28.2851              Number of Variables   :        3
S.D. dependent var  :      2.7671              Degrees of Freedom    :       44
R-squared           :      0.9243
Adjusted R-squared  :      0.9208
Sum squared residual:      26.667              F-statistic           :  268.5771
Sigma-square        :       0.606              Prob(F-statistic)     : 2.196e-25
S.E. of regression  :       0.779              Log likelihood        :   -53.372
Sigma-square ML     :       0.567              Akaike info criterion :   112.744
S.E of regression ML:      0.7532              Schwarz criterion     :   118.295

-----------------------------------------------------------------------------------
            Variable    Coefficient        Std.Error     t-Statistic    Probability
-----------------------------------------------------------------------------------
            CONSTANT     19.1830728        1.9241611       9.9695774      0.0000000
           birthrate     -0.7761151        0.1580456      -4.9107020      0.0000129
           deathrate      1.3461441        0.0837745      16.0686685      0.0000000
-----------------------------------------------------------------------------------

REGRESSION DIAGNOSTICS
MULTICOLLINEARITY CONDITION NUMBER          36.021

TEST ON NORMALITY OF ERRORS
TEST                            DF         VALUE           PROB
Jarque-Bera                      2           1.412           0.4937

DIAGNOSTICS FOR HETEROSKEDASTICITY
RANDOM COEFFICIENTS
TEST                            DF         VALUE           PROB
Breusch-Pagan test               2           0.393           0.8215
Koenker-Bassett test             2           0.553           0.7583

SPECIFICATION ROBUST TEST
Not computed due to multicollinearity.

DIAGNOSTICS FOR SPATIAL DEPENDENCE
TEST                          MI/DF        VALUE           PROB
Moran's I (error)            0.3948          6.989           0.0000
Lagrange Multiplier (lag)         1          3.271           0.0705
Robust LM (lag)                   1          0.012           0.9116
Lagrange Multiplier (error)       1         27.817           0.0000
Robust LM (error)                 1         24.559           0.0000
Lagrange Multiplier (SARMA)       2         27.830           0.0000
```

Figure 32 Model with Birth rate and Death rate

Model using birth rate, death rate, ratio of DID

```
REGRESSION
----------
SUMMARY OF OUTPUT: ORDINARY LEAST SQUARES
-----------------------------------------
Data set            :finalmodel_var3.dbf
Weights matrix      :File: SRA_DB_JPN_knn7.gwt
Dependent Variable  :  aged_ratio              Number of Observations:      47
Mean dependent var  :     28.2851              Number of Variables   :       4
S.D. dependent var  :      2.7671              Degrees of Freedom    :      43
R-squared           :      0.9255
Adjusted R-squared  :      0.9203
Sum squared residual:      26.230              F-statistic           :   178.1366
Sigma-square        :       0.610              Prob(F-statistic)     :  2.869e-24
S.E. of regression  :       0.781              Log likelihood        :    -52.984
Sigma-square ML     :       0.558              Akaike info criterion :    113.968
S.E of regression ML:      0.7471              Schwarz criterion     :    121.368

-----------------------------------------------------------------------------------
            Variable     Coefficient       Std.Error     t-Statistic    Probability
-----------------------------------------------------------------------------------
            CONSTANT      17.5036178       2.7682147       6.3230709      0.0000001
           birthrate      -0.7377354       0.1649128      -4.4734866      0.0000556
           deathrate       1.4301666       0.1300648      10.9957987      0.0000000
             did_pop       0.0083004       0.0098060       0.8464618      0.4019815
-----------------------------------------------------------------------------------

REGRESSION DIAGNOSTICS
MULTICOLLINEARITY CONDITION NUMBER           56.744

TEST ON NORMALITY OF ERRORS
TEST                            DF         VALUE           PROB
Jarque-Bera                      2          1.191           0.5513

DIAGNOSTICS FOR HETEROSKEDASTICITY
RANDOM COEFFICIENTS
TEST                            DF         VALUE           PROB
Breusch-Pagan test               3          0.519           0.9147
Koenker-Bassett test             3          0.663           0.8819

SPECIFICATION ROBUST TEST
Not computed due to multicollinearity.

DIAGNOSTICS FOR SPATIAL DEPENDENCE
TEST                          MI/DF        VALUE           PROB
Moran's I (error)            0.3860         6.967           0.0000
Lagrange Multiplier (lag)        1          2.425           0.1194
Robust LM (lag)                  1          0.098           0.7539
Lagrange Multiplier (error)      1         26.584           0.0000
Robust LM (error)                1         24.257           0.0000
Lagrange Multiplier (SARMA)      2         26.682           0.0000
```

Figure 33 Model with birthrate, death rate and DID

Model using Birthrate, death rate, fertility

```
REGRESSION
----------
SUMMARY OF OUTPUT: ORDINARY LEAST SQUARES
-----------------------------------------
Data set            :finalmodel_var3.dbf
Weights matrix      :File: SRA_DB_JPN_knn7.gwt
Dependent Variable  :  aged_ratio            Number of Observations:         47
Mean dependent var  :    28.2851             Number of Variables   :          4
S.D. dependent var  :     2.7671             Degrees of Freedom    :         43
R-squared           :     0.9282
Adjusted R-squared  :     0.9232
Sum squared residual:     25.278             F-statistic           :   185.3832
Sigma-square        :      0.588             Prob(F-statistic)     :  1.298e-24
S.E. of regression  :      0.767             Log likelihood        :    -52.115
Sigma-square ML     :      0.538             Akaike info criterion :    112.231
S.E of regression ML:     0.7334             Schwarz criterion     :    119.631

-----------------------------------------------------------------------------------
            Variable     Coefficient       Std.Error     t-Statistic     Probability
-----------------------------------------------------------------------------------
            CONSTANT      20.5078171       2.0818363       9.8508308       0.0000000
           birthrate      -1.2800954       0.3629623      -3.5267995       0.0010142
           deathrate       1.1470088       0.1535994       7.4675332       0.0000000
           fertility       3.1606614       2.0563443       1.5370293       0.1316127
-----------------------------------------------------------------------------------

REGRESSION DIAGNOSTICS
MULTICOLLINEARITY CONDITION NUMBER          82.551

TEST ON NORMALITY OF ERRORS
TEST                              DF         VALUE           PROB
Jarque-Bera                        2           1.195           0.5503

DIAGNOSTICS FOR HETEROSKEDASTICITY
RANDOM COEFFICIENTS
TEST                              DF         VALUE           PROB
Breusch-Pagan test                 3           0.299           0.9602
Koenker-Bassett test               3           0.374           0.9455

SPECIFICATION ROBUST TEST
Not computed due to multicollinearity.

DIAGNOSTICS FOR SPATIAL DEPENDENCE
TEST                            MI/DF        VALUE           PROB
Moran's I (error)               0.3679         6.755           0.0000
Lagrange Multiplier (lag)          1           8.197           0.0042
Robust LM (lag)                    1           1.340           0.2470
Lagrange Multiplier (error)        1          24.156           0.0000
Robust LM (error)                  1          17.300           0.0000
Lagrange Multiplier (SARMA)        2          25.497           0.0000
```

Figure 34 Model with Birthrate, death rate and fertility

Model using only Fertility. Results in spatially lagged model

```
REGRESSION
----------
SUMMARY OF OUTPUT: ORDINARY LEAST SQUARES
-----------------------------------------
Data set            :finalmodel_var3.dbf
Weights matrix      :File: SRA_DB_JPN_knn7.gwt
Dependent Variable  :  aged_ratio              Number of Observations:         47
Mean dependent var  :     28.2851              Number of Variables   :          2
S.D. dependent var  :      2.7671              Degrees of Freedom    :         45
R-squared           :      0.0012
Adjusted R-squared  :     -0.0210
Sum squared residual:     351.802              F-statistic           :     0.0535
Sigma-square        :       7.818              Prob(F-statistic)     :     0.8182
S.E. of regression  :       2.796              Log likelihood        :   -113.994
Sigma-square ML     :       7.485              Akaike info criterion :    231.987
S.E of regression ML:      2.7359              Schwarz criterion     :    235.688

-----------------------------------------------------------------------------
            Variable    Coefficient       Std.Error     t-Statistic    Probability
-----------------------------------------------------------------------------
            CONSTANT     27.1709246       4.8362927       5.6181307      0.0000011
           fertility      0.7284260       3.1505922       0.2312029      0.8182052
-----------------------------------------------------------------------------

REGRESSION DIAGNOSTICS
MULTICOLLINEARITY CONDITION NUMBER          23.674

TEST ON NORMALITY OF ERRORS
TEST                             DF          VALUE           PROB
Jarque-Bera                       2          6.887          0.0320

DIAGNOSTICS FOR HETEROSKEDASTICITY
RANDOM COEFFICIENTS
TEST                             DF          VALUE           PROB
Breusch-Pagan test                1          7.136          0.0076
Koenker-Bassett test              1          4.544          0.0330

SPECIFICATION ROBUST TEST
TEST                             DF          VALUE           PROB
White                             2         30.121          0.0000

DIAGNOSTICS FOR SPATIAL DEPENDENCE
TEST                           MI/DF         VALUE           PROB
Moran's I (error)             0.1916         3.446          0.0006
Lagrange Multiplier (lag)         1          6.701          0.0096
Robust LM (lag)                   1          1.190          0.2753
Lagrange Multiplier (error)       1          6.550          0.0105
Robust LM (error)                 1          1.039          0.3080
Lagrange Multiplier (SARMA)       2          7.740          0.0209
```

Figure 35 Model with Fertility

Model using DID and fertility

```
REGRESSION
----------
SUMMARY OF OUTPUT: ORDINARY LEAST SQUARES
-----------------------------------------
Data set            :finalmodel_var3.dbf
Weights matrix      :File: SRA_DB_JPN_knn7.gwt
Dependent Variable  :  aged_ratio              Number of Observations:         47
Mean dependent var  :    28.2851               Number of Variables   :          3
S.D. dependent var  :     2.7671               Degrees of Freedom    :         44
R-squared           :     0.5337
Adjusted R-squared  :     0.5125
Sum squared residual:    164.231               F-statistic           :    25.1826
Sigma-square        :      3.733               Prob(F-statistic)     :   5.13e-08
S.E. of regression  :      1.932               Log likelihood        :    -96.092
Sigma-square ML     :      3.494               Akaike info criterion :    198.183
S.E of regression ML:     1.8693               Schwarz criterion     :    203.734

-----------------------------------------------------------------------------------
            Variable    Coefficient       Std.Error     t-Statistic     Probability
-----------------------------------------------------------------------------------
            CONSTANT     46.1725715       4.2839226      10.7781060       0.0000000
             did_pop     -0.1210642       0.0170779      -7.0889504       0.0000000
           fertility     -7.5363910       2.4694965      -3.0517926       0.0038483
-----------------------------------------------------------------------------------

REGRESSION DIAGNOSTICS
MULTICOLLINEARITY CONDITION NUMBER         34.612

TEST ON NORMALITY OF ERRORS
TEST                             DF         VALUE            PROB
Jarque-Bera                       2          1.187           0.5524

DIAGNOSTICS FOR HETEROSKEDASTICITY
RANDOM COEFFICIENTS
TEST                             DF         VALUE            PROB
Breusch-Pagan test                2          1.416           0.4925
Koenker-Bassett test              2          1.879           0.3907

SPECIFICATION ROBUST TEST
Not computed due to multicollinearity.

DIAGNOSTICS FOR SPATIAL DEPENDENCE
TEST                           MI/DF        VALUE            PROB
Moran's I (error)              0.3144        5.417           0.0000
Lagrange Multiplier (lag)         1         18.778           0.0000
Robust LM (lag)                   1          2.931           0.0869
Lagrange Multiplier (error)       1         17.634           0.0000
Robust LM (error)                 1          1.787           0.1813
Lagrange Multiplier (SARMA)       2         20.565           0.0000
```

Figure 36 Model with DID and Fertility

Model using Jobs taken out of prefecture, migration, and secondary industry

```
REGRESSION
----------
SUMMARY OF OUTPUT: ORDINARY LEAST SQUARES
-----------------------------------------
Data set            :finalmodel_var3.dbf
Weights matrix      :File: SRA_DB_JPN_knn7.gwt
Dependent Variable  :  aged_ratio              Number of Observations:        47
Mean dependent var  :    28.2851               Number of Variables   :         4
S.D. dependent var  :     2.7671               Degrees of Freedom    :        43
R-squared           :     0.5476
Adjusted R-squared  :     0.5160
Sum squared residual:    159.352               F-statistic           :    17.3481
Sigma-square        :      3.706               Prob(F-statistic)     :  1.575e-07
S.E. of regression  :      1.925               Log likelihood        :    -95.383
Sigma-square ML     :      3.390               Akaike info criterion :    198.766
S.E of regression ML:     1.8413               Schwarz criterion     :    206.166

-----------------------------------------------------------------------------------
            Variable    Coefficient       Std.Error    t-Statistic    Probability
-----------------------------------------------------------------------------------
            CONSTANT     29.4560109       2.3523472     12.5219657      0.0000000
             jobsout     -0.0465909       0.0362219     -1.2862648      0.2052335
           migration    -11.6286482       1.8587008     -6.2563313      0.0000002
           secondind     -0.0837907       0.0784924     -1.0675009      0.2917025
-----------------------------------------------------------------------------------

REGRESSION DIAGNOSTICS
MULTICOLLINEARITY CONDITION NUMBER          20.305

TEST ON NORMALITY OF ERRORS
TEST                             DF        VALUE           PROB
Jarque-Bera                       2         34.993          0.0000

DIAGNOSTICS FOR HETEROSKEDASTICITY
RANDOM COEFFICIENTS
TEST                             DF        VALUE           PROB
Breusch-Pagan test                3          8.878          0.0310
Koenker-Bassett test              3          3.129          0.3722

SPECIFICATION ROBUST TEST
TEST                             DF        VALUE           PROB
White                             9         27.240          0.0013

DIAGNOSTICS FOR SPATIAL DEPENDENCE
TEST                           MI/DF       VALUE           PROB
Moran's I (error)              0.0719        1.767          0.0773
Lagrange Multiplier (lag)         1          2.343          0.1258
Robust LM (lag)                   1          1.457          0.2275
Lagrange Multiplier (error)       1          0.923          0.3368
Robust LM (error)                 1          0.036          0.8499
Lagrange Multiplier (SARMA)       2          2.379          0.3044
```

Figure 37 Model with jobs, migration, and secondary industry