# Final Paper

*Alice Mee Seon Chung and Ningyin Xu*

## I. Introduction

Since China has been one of key players of world economy, many studies have been conducted on the economic development and income level of China. The income level of people directly related with the economic power so it is important to know the prediction on how this income level will change in the future at various level like local, community, family and individual. In this project, our goal is to study what factors affect the level of individual income in China. It is commonly believed that the socioeconomic backgrounds and characteristics of residential community have large impact on the income level of individual. To extent the existed findings, we proposed the research question on what factors impact personal income and how they impact. Our main data source is China Family Panel Studies (CFPS), the first large-scale academically oriented longitudinal survey data in China. We used data for 2 periods, with 33,018 observations in 2010 and 37,147 observations in 2014. After cleaning the missing variables, selecting the interested variables and merging two data by personal id, we ended up with 12,237 observations and 202 variables. We focused on the important variables associated with personal income: age, gender, years of education and other variables on community characteristics: the number of convenience stores, the number of primary schools, the number of hospitals, water source, community population, travel time to nearest town, travel time to county seat, agricultural GDP, non-agricultural GDP and income per capita.

Using these variables, we can measure personal mean income by each province and further inspect the relationship of it with individual and community characteristics. First, we will examine the overall features of these variables to get the sense of understanding the data. Then, we will run a multivariate regression with the variables and examine the prediction of model.

## II. Graphical Methods Methods

To answer our research question, the selection of data visualization methods is significant and the most important part of our research to deliver meaningful information of the data. As we have province id in the dataset, we can look personal income by each province. Based on the characteristics of our data, we can

use two different types of graph: geospatial and statistical. Technically, we mainly use the graphic package ggplot2 in r and plot function in package lmtest.

Since we focus on what factors impact on the income level, we need to investigate the relationship between the dependent variable, income, and the independent variables. We examine the interested variables considering their specific characteristics and type of variable, then choose the effective graphical method to show its distribution and the relationship with personal income.

First, our main dependent variable is personal income so we start to explore to see the relationship between the mean of personal income and GDP in each province.

*Province*

In the first plot, "Mean of Personal Income vs. GDP", we use scatter plot with geom_point to show the relationship of two variables, mean of personal income and GDP of each province in China. Here, we intend to provide overall distribution and relationship of two variables without giving any information on which dot is corresponding to which province in China. We use color to indicate year and also use title, x-axis, y-axis, legend, and captions to convey the purpose and information to interpret the plot. From this plot we can see that mean of personal income and GDP in each province has positive relationship and both of two variables increased in 2014. From above plot, we can question that how are the difference of mean personal income between 2010 and 2014 across China. We use geospatial visualization to show this distribution using China map. Since R studio does not have data to draw China map with provincial geospatial data, longitude and latitude, we referenced the ggplot2-china-map in github repository (https://github.com/xiaohk/ggplot2-china-map) to draw each province in China.

For the second plot, "Difference of Mean Personal Income between 2014 and 2010 in China", we use geom_polygon and geom_path to draw the whole China map and set the fill in geom_polygon and use scale_fill_gradient to show the difference scale with color in ggplot2. We use the color scale to indicate the level of difference, so the darker means higher difference in mean personal income. To make consistent color palette, we select the turquoise color and also use title and captions to deliver the additional information like grey area means there are no available income information in our data. From this plot we can see overall distribution of difference in mean personal income and perceive which area has the highest difference.

Further, in the third and fourth plot, "Difference of Mean Personal Income between 2014 and 2010 in East China" and "Difference of Mean Personal Income between 2014 and 2010 in Midwest China", we divide regions into two group, east and midwest, to see closely which area has become richer than the other in four years. We use subset of provinces in China and manually set the scale be the same across the three geospatial

plots to make easy to compare. Using same color and scale bound, we can observe that east region in China has become richer than midwest region in China in four years.

Next, we move to examine the distribution of interested variables that we stated above and the relationship of those and personal income. These variables also categorize as individual level and community level. We first look the variables related with the characteristic of individuals and then move to the variables related with the characteristics of residential communities. Here, we use statistical visualization method using ggplot2 and apply specific graphical functions considering the type of variable. For every variable, we first look at the distribution and then examine the relationship between the variable and personal income. For basic setting to draw the graphs, we use consistent color palette used in the previous plots and the same colors to indicate the same year as well.

*Individual Level*

*Age*

For the distribution, we use geom_line with stat setting as density. We can easily see how the distribution looks like and compare the data with two years using colors. For the relationship, since age is discrete variable, we can not draw the line graph directly so we use geom_smooth with span above the scatter plot using geom_point. From the plot, we can see that until age 30, it has positive relationship but after that, it turns to have negative relationship.

*Age-squared*

For the distribution, we use the same ggplot2 method as in the age variable since the characteristic of two variables are similar. To see the relationship, we choose to draw geom_smooth with method as "lm" and span above the scatter plot using geom_point since age-squared is also discrete variable. The relationship between age-squared and personal seems to be negative from the plot.

*Gender*

To show the distribution, we use geom_bar with position as dodge since the gender is categorical variable. It is useful to use bar chart to convey information and show the distribution when the variable is categorical. For the relationship, since gender is categorical variable, we can not directly compare the relationship. So we use geom_boxplot with position as dodge to compare the relationship between the gender and personal income. From the plot, we can compare between the gender and male seems to have higher personal income than female in both years.

*Education Years*

We use geom_line with stat setting as density since education is discrete variable for the distribution. The overall distribution looks similar within two years, but the fluctuation of density in 2010 is larger than 2014. In the part of showing relationship, we can not draw the line graph directly since education years is discrete variable. So we use we choose to draw geom_smooth with method as "lm" and span above the scatter plot using geom_point. From the plot, we can see that as education years get larger personal income also gets larger. So we can observe the positive relationship between two variables and the increasing trend is steeper in 2014 than 2010.

*Residential Community Level*

*The Number of Convenience Stores in Communities*

For the distribution, we use geom_line with stat setting as density since the number of convenience stores is discrete variable. The overall distribution looks a lot similar within two years and this variable has large variance. For the relationship, we can not draw the line graph directly so we use we choose to draw geom_smooth with meodel as "lm", linear moel above the scatter plot using geom_point. From the plot, we can see that as the number of convenience stores gets larger, personal income also gets larger. So we observe the positive relationship between two variables and this relationship is stronger in 2010 than 2014 from the plot.

*The Number of Primary Schools in Communities*

To see the distribution, we use geom_bar with position as dodge because the number of primary schools is discrete variables with small number and the difference of count between those numbers is large. In this case, using bar chart is more suitable to show the distribution. For showing relationship, we use geom_boxplot with position as dodge to compare the relationship between the two variables and resize the box width to indicate the group size. From the plot, there is positive relationship in 2010, while we can not tell there is any strong relationship in 2014.

*Types of Water Sources in Communities*

We use geom_bar with position as dodge since types of water sources are categorical variables for the distribution part. We can not directly compare the relationship beacuse types of water source are categorical variables in the relationship plot part. We again use geom_boxplot with position as dodge to compare the relationship between two variables and resize the box width to indicate the size of group as we did with the number of primary schools. From the plot, well spring water and tap water seem to have relationship with personal income in 2010 and tap water and other source seem to have positive relationship with personal

income in 2014.

*The Number of Hospitals in Communities*

For the distribution, we use geom_line with stat setting as density since the number of convenience stores is discrete variable. The distribution is similar in both 2010 and 2014. For the relationship, we use geom_boxplot with position as dodge to compare the relationship between the two variables. From the plot, overall we can see there is positive relationship between the number of hospitals and personal income in both years.

*Population in Communities*

We use geom_line with stat setting as density since population is discrete variable to show the distribution. The overall distributions look alike with slightly higher density in larger population in 2014. This variable also has large variance. We choose to use geom_smooth with method as "lm" over the scatter plot with geom_point to show relationship of two variables. From the plot, we can see that as the population gets larger, personal income also gets larger. So we can see the positive relationship between two variables in general.

*Travel Time to Nearest Town in Communities*

For the distribution, travel time to nearest town is continuous variable so we use geom_line with stat setting as density. In 2010, the density is condensed around 1 to 2 hours. For the relationship, we choose to draw geom_smooth with method as "lm" and span above the scatter plot using geom_point. From plot, we can see there is negative relationship between two variables and 2014 has large bound on travel time than 2010.

*Travel Time from County Seat in Communities*

For presenting the distribution, travel time from county seat is continuous variable, so we decide to use geom_line with stat setting as density to show the distribution. The density is centered at less than 5 hours in both years. For the relationship, we use geom_smooth with method as "lm" and span above the scatter plot using geom_point as we did with previous variable. We can see there is negative relationship between two variables in 2010 while we can not tell there is nay relationship in 2014 in the plot.

*Agricultural GDP in Communities*

For the distribution, we again use geom_line with stat setting as density to present the distribution. We can see how the distribution looks like and compare the data with two years by colors and this variable also has large variance. To present the relationship of two variables, so we use geom_smooth with span above the scatter plot using geom_point since agricultural GDP in communities is discrete variable and hard to draw

the relationship as line. From the plot, we can see several curves around GDP under 2500 so it is hard to tell the relationship is explicitly positive or negative.

*Non-Agricultural GDP in Communities*

The type of non-agricultural GDP is discrete variable so we use geom_line with stat setting as density to show the distribution. We use color to distinguish two years and from the plot, we can see that the distribution of non-agricultural GDP in 2014 has large variance than in 2010. We draw geom_quantile with method as rqss and with four-quantiles interval to see the relationship. From the plot as non-agricultural GDP gets larger, personal income also gets larger in all four quantiles and two years. Thus, there is positive relationship between two variables.

*Income per capita in Communities*

As the type of Income per capita is discrete variable, so we use geom_line with stat setting as density to show the distribution of this variable. In the plot, income per capita in 2010 shows higher density under 2500. 2014 has higher density when income per capita is larger than approximately 5500. Overall, income per capita in 2014 is larger than in 2010. To draw the relationship, we use geom_smooth and span over the scatter plot using geom_point in this part. We can observe that income per capita and personal income has positive relationship in general.

After we examine the characteristics of all interested variables, we now can have the sense that how and why include all these variables to see how personal income changes. With these variables, we ran the regression model with panel data and it is defined as below:

$$log(income) = \beta_0 + \beta_1 age + \beta_2 age^2 + \beta_3 educationyear + \beta_4 numberofconveniencestore$$

$$+\beta_5 numberofprimaryschool + \beta_6 numberofhospitals + \beta_7 watersource + \beta_8 pupulation$$

$$+\beta_9 timecommittonearesttown + \beta_{10} timecommittocountyseat + \beta_{11} agriculturalGDP$$

$$+\beta_{12} Non - agriculturalGDP + \beta_{13} incomepercapita + \beta_{14} year$$

To show and validate of the model, we plot the correlation plot with selected numerical variable and coefficient plot from the model.

To validate the independence of variables, we decide to plot correlation plot. We use package corrplot to plot the correlation plot using continuous and discrete variables. We omit the categorical variable when plotting

the correlation plot since those are not suitable for testing the independence. Using the result of data frame of correlation scores within selected variables, we can draw correlation plot. The range is from -1 to 1 and the score close to 1 or -1 means they are correlated each other. In the plot, we use the diverging color scale using two consistent colors that we keep suing through the whole graphs. The lighter color means to get close to 0 and those are not correlated and independent each other. The correlation plot shows that our variables are independent enough to say they are not correlated so that we can further use those variables to run the model.

We use package coefplot to plot the coefficient of all variables. Coefficient plot is useful graphical tool to compare the estimate of all coefficients and we also can show the estimate with confidence interval. We use 95% confidence interval in this plot. We use consistent color palette on this plot and do not apply variety of colors to every estimates to avoid the confusion. The graph itself can clearly deliver the information. Most of the variables except Year 2014, Tap water, and Number of hospitals in communities are close to 0 or have positive relationship with personal income.

## III. Result and Discussion

From the work and graphs, the audience has thoroughly learned what and how factors impact on personal income in China. The audience examine the main variable, personal income, and the factors in two aspects: geospatial and statistical. Using two approach, we present the distribution of difference in personal income by province and the characteristics of each variables. From these visualizations, the audience can learn and understand the caveats of personal income in China. Additionally, we present the linear model using those variables and how the variables impact on personal income in China. We also validate the variables using correlation plot and visualize the coefficients of model to help audience to understand. The audience could learn how the factors impact on personal income from this part.

*Truthful*

The visualization is truthful since we use the actual survey-based data from China in 2010 and 2014, which is the first large-scale academically oriented longitudinal survey data in China. We tried to include all validate observation and cleaned the data using personal id to identify the unique observations across two years. We tried to add information using title, subtitle, and notation if we apply additional exclusion to draw graph or if there are anything that need to be clarified such as unit or color.

*Functional*

We believe this visualization is functional since we tried to use the visualization techniques considering the flow to understand our research question and dataset with several steps start from general concept to specific variables. We select the graphic types to support the character tics of variables and to show findings effectively. We tried to arrange all the graphs in a logical way: 1) general scatter plot of personal income and GDP in China; 2) geographical map of the difference in personal income in China; 3) statistical graph to describe the distribution of each interested variable; 4) statistical graph to show the relationship of personal income and each interested variable; 5) statistical graph to validate and show the result of the linear regression model using all interested variable. The audience can follow this flow and understand the story. Moving forward, the audience keep accumulating the characteristic and overall features of the variables to understand the findings. The graph itself is understandable to the audience as we added some texts and notations on the graphs.

*Beautiful*

We mainly use red and green as two main colors. As we tried to be consistent on using color palette in all graphs, we believe the graphs are clear, simple and very explicitly presenting the information to the audience. The main colors corresponding to each year, 2010 and 2014, so the color helps to interpret the graphs of variables. For presenting the difference in personal income growth in the maps, we use the diverging scale of color to indicate the importance. Similar diverging scale with two main colors is used in the correlation plot. The graphs are aesthetically pleasing, attractive and intriguing because of simplicity and clarity in the plots.

*Insightful*

The graphs are insightful to provide and reveal the characteristics of variables and the findings of our research question. The insight is built as the audience follows the logical chain of graphs and understand the meaning of graphs. We try to provide both "a-ha" moment and gradual process of examination of the story on research question using this logical chain. At first, scatter plot gives the curiosity of personal income growth in China from 2010 to 2014. Then, we show the personal income difference with China Map and additional regional comparison between East China and Midwest China. Next, we try to present the specific features of each variables and the relationship with personal income. From the insights that gained from previous graphs, the audience can understand and interpret the result of linear regression model in the final part.

*Enlightening*

Taken all together of four qualities and the findings of our research question, the visualization is enlightening to the researcher who wants to study of this dataset. Since this dataset contains lots of variables so one may feel lost when they first look at the data. The visualization is enlightening in the sense that we provide the way to present the information with logical chain and apply several different graphical techniques. We apply

geospatial visualization to present the income difference between two years instead of drawing a line graph. Instead of showing the result table of regression model, we visualize the result to make easy to interpret. The story and conclusion from the visualization could provide insight or implication to the policy makers or government to boost personal income in future China.