sustainability

MDPI

Article

Internal Differentiation within the Rural Migrant Population from the Sustainable Urban Development Perspective: Evidence from China

Xiaohong Deng, Lei Gong *, Yanfang Gao, Xiaoqing Cui and Ke Xu

School of Management Engineering, Shandong Jianzhu University, 1000 Fengming Road, Licheng District, Jinan 250101, China; xhdeng@sdjzu.edu.cn (X.D.); gaoyanfang@sdjzu.edu.cn (Y.G.); cxq6706@sdjzu.edu.cn (X.C.); xukeooop@163.com (K.X.)

* Correspondence: gonglei1230@163.com

Received: 14 November 2018; Accepted: 17 December 2018; Published: 18 December 2018

Abstract: Population mobility and attendant issues, especially housing issues, have a major impact on sustainable urban development. In the urbanization process, a number of micro-communities with various social characteristics have come to compose the rural migrant population (RMP), resulting in internal differentiation. This study aims to reveal the demographic structure of this specific group, and to analyze the effects of the mechanism between population flow trends and sustainable urban development, taking housing demand as a starting point. To this end, a clustering model for mixed-type data based on partitioning around the medoid is proposed, and the linked characteristics and potential laws of the RMP are analyzed, based on the dynamic data of the migrant population in eastern China. To achieve sustainable urban development, the locational preferences and coping strategies of inflowing micro-communities based on city types are demonstrated. The results show that the RMP can be divided into four groups that have strong representativeness and that show significant differences in population structure and housing demand. Super-large and medium-sized cities are the main migration destinations. Several suggestions are proposed, based on these results. Housing security policies should be designed according to the housing demand characteristics and the spatial distribution of different groups. Housing security policies should play a full and positive role in reasonably guiding RMP movement.

Keywords: rural migrant population; internal differentiation; sustainable urban development; cluster analysis; eastern China

1. Introduction

There is a special migrant group called the rural migrant population (RMP) in China, because of the dual household registration system [1]. Since the reform and opening-up, more rural surplus labor has transferred to cities in the accelerated urbanization process in China, and this has become the main body of industrial workers [2–4]. According to the National Bureau of Statistics, the number of peasant workers in China has reached 286.52 million, among which 171.85 million or more than 70.43% are migrant workers [5]. Due to the country's large population base and rapid growth, the conversion of the RMP from farmers to residents must be a step-by-step process. In this process, there is differentiation by social class within the RMP, which generates a series of micro-communities with various characteristics. The development trends between these micro-communities manifest differently, according to aspects such as social factors. As a result, various demands occur for social resources, particularly housing demand, and these demands must receive attention as a critical part of sustainable urban development [6].

On the other hand, the RMP is not eligible for the social benefits resulting from economic growth because of household registration restrictions [7,8]. This population's position is on the edge of urban society; that is, it is excluded from urban society [6,9]. Although this social status quo has been improved by the housing security policy, current policy does not perfectly match the population's needs [10]. A serious contradiction between the social status quo and actual demand is generated, which poses a severe challenge to the sustainable urban development of population and housing [11].

In the field of RMP, previous studies related to the complex relation between housing demand and housing security patterns have contributed to discovering the significance of improving housing policies based on actual demand. However, most researchers analyze this issue using theoretical logical analysis or case studies from the perspective of the housing or the security status quo of the RMP [1,12,13]. This might be because of the difficulty of collecting data related to the RMP. There is less empirical analysis in this research field, based on statistical data from systematic surveys, and few studies associate the heterogeneity of the RMP with the housing demand.

The major objectives of this paper are (1) to analyze the actual housing demand of the RMP from the perspective of internal differentiation, and (2) to put forward proposals based on the improvement of housing security policies to promote the sustainable development of urban housing. To achieve these goals, we propose an improved clustering analysis algorithm to demographically differentiate groups within the RMP, which contributes to detecting the defects of the macroeconomic housing policies in China. To make the results more persuasive, our empirical analysis is discussed based on the microdata from the China Migrants Dynamic Survey (CMDS), which is the novelty of this study.

This paper is structured as follows. Section 2 introduces the previous research and background of the internal differentiation within the RMP and the population's housing demands. Section 3 presents the data sources and definitions of influencing factors. Section 4 elaborates the principle of the data processing algorithm used. Section 5 outlines the experimental process and discusses the analytical results. Finally, the major conclusions are summarized, based on which the corresponding policy suggestions are given.

2. Literature Review

2.1. Internal Differentiation of Migrants

Academically, rural migrants have been systematically defined by scholars as populations that float from rural to urban areas [14]. Non-agricultural workers or surplus labor from rural areas (the so-called "peasant workers") represent the object we focus on in this paper. The serious social issues caused by this special population have long been a focus of attention in academic circles. Some researchers have developed concepts and correlative theories to uncover the social roots and inherent development characteristics of this heterogenous population. However, the majority of studies focus on sociology and demography. One study found significant social class differences within RMP groups, in what is known as social stratification, and the initial understanding of the RMP as a whole was gradually displaced [15,16]. Another study found that more distinctions between individual and sometimes micro-communities are coming to be universally acknowledged [17]. Essentially, this conclusion indicates that the groups within the RMP are not homogeneous. As a result of internal differentiation, the RMP can be generally divided into many contrasting micro communities, also called clusters, by using certain criteria (see Section 2.2) [18].

Another research method is to analyze micro survey data by statistical methods to explore the population classification and its related social problems. Previous studies with different aims have discussed this issue from multiple perspectives, differentiating the population by considering aspects such as gender [19], race [20], social services [18], generation [21], occupation [22], and urban identity [23]. The internal differentiation of migrants may be a reason for the demand for different social recourses. Sustainable housing is the fundamental material prerequisite for living, as well as

a key factor in livelihood issues [24]. Exploiting the internal connection between differentiation and housing demand is becoming a new research hotspot in the field of sociology.

2.2. Influencing Factors of Housing Demand

The social problems of RMP are regional, which has led to little academic research worldwide, with the exception of China [25]. However, there are abundant publications in the field of housing demand, which can be referenced in our study. Previous research on this issue consists of two aspects: macroscopic analysis and microcosmic analysis. From the macroscopic perspective, the connection between statistical indicators and sustainable supply and demand in the housing market is discussed based on the national or social level. Kagochi and Mace examine the influencing factors of housing demand for a single household, based on panel time series data for 1988–2007, and elaborate the results, including population growth, sale of existing houses, cost of building, and unemployment rates [26]. Similarly, Turkish scholars report that real mortgage interest rates and household expenditure are regarded as the major demand-related factors, on the basis of a survey of recent housing projects from 2016 to 2018 [27]. In addition, some researchers have found that the housing loan rate and inflation have a significant effect on housing price, thereby further affecting housing demand [28]. The macroscopic analysis method is sufficient to identify the demand changes of the real estate market, but it is unable to reflect the individual housing demand. To solve this problem, through microdata, the housing demand of a special population group is proposed to obtain the specific requirements on the individual level.

Due to the immense diversity of the datasets used for analysis, previous studies yield significantly different final conclusions. Research on housing demand may have started with Mankiw and Weil's empirical analysis. Their results were obtained by a statistical regression model applied to American census data for 1970, which is regarded as a powerful tool for describing the reasons for changes in the housing price and housing demand. The intergenerational difference, namely, age, has been shown to be an influencing factor of housing demand by many researchers [29–33]. With the improvement of the model and dataset, more potential confounders have gradually been identified. Han presented a particular method to estimate the housing preference, which can simulate the actual housing demand of Shanghai residents in assumed conditions that are related to household income, family size and generation. The price–income ratio, age composition, and household size were found to have significant effects on housing demand [30]. In addition to household size and income, career has been found to be a potential stimulus for housing demand. This argument was supported by Oktay by using survey data of households living in the Erzurum city centre [34]. Eichholtz studied the housing conditions of English residents to describe the connection between demographic characteristics and the demand for residential real estate, and found that housing demand is influenced by the price-income ratio, age composition, and household size [31]. Flambard estimated the housing demand in northern France according to sociodemographic factors, and noted that residential choices strongly depend on the rent-to-income ratio and the distance to work. Moreover, housing preferences display striking differences in different locations [32]; we elaborate on this issue in Section 5.2. A recent study related to rural–urban migration in China evaluated the current dynamics of living, and creatively showed that an individual's native place might have an enormous impact on the housing demand of Chinese migrants. The idea that improving the housing system is an effective approach to resolving the housing problem for migrants in China is elaborated at the end of this paper [6]. Through the above expositions, we find that there are great differences in the influencing factors of regional research, although some factors, such as age, income, education, and household size, are widely recognized. More potential influencing factors in addition to these common factors should be identified according to local characteristics. Such an analysis can clearly reflect the demand for housing, which is the first and most critical step in promoting the sustainable development of urban housing.

3. Data and Feature Selection

3.1. The Data

Our empirical analysis of the internal differentiation of the RMP is derived from the microdata on the RMP in the CMDS [35], a continuous research project that is designed by the National Health and Family Planning Commission of China, and is implemented by the Population and Development Research Center. The survey collected valuable information on the RMP and its households, including demographic characteristics, housing attributes, income structure, and consumer status. The subjects of the research area were individuals belonging to the RMP (aged 19 to 59) who had lived in the area for more than one month.

We selected the CMDS as our basic dataset for three reasons. First, the CMDS dataset covers up to 320 dimensions, which is enough for the empirical analysis. Second, the CMDS adopts a stratified three-stage probability that is proportionate to size sampling, which is the most credible census method. Third, the CMDS removes temporary floating populations, such as students and visiting relatives. The screening of the migrant population by the CMDS was consistent with the data conditions demanded in this paper, and provided sufficient screening space for the empirical analysis.

We collected the latest CMDS data for 2014 as the sample of empirical analysis. The CMDS 2014 dataset consisted of 82,985 valid data items collected from 10 cities in eastern China. According to the definition of the RMP by the National Bureau of Statistics and this paper, the qualified data should meet the following requirements:

1. The value of *simp_type* (sample type) must be the residents' committee, which means these respondents lived in urban areas.
2. The value of *acc_nat_1* (account nature) must be an agricultural registered permanent residence
3. The value of *flo_reason_1* (flow reason) must be to do business or to seek a job in the city.

We obtained a preliminary dataset consisting of 36,764 RMP samples on the basis of the above principles. Then, we filtered out 1260 samples with either missing critical values or logical errors from the preliminary dataset. Our final dataset consisted of 35,504 samples.

3.2. Feature Selection and Descriptive Statistics

There were 320 variables related to individuals and households in the original dataset, some of which were redundant for this study, and had to be eliminated. Due to the insufficiency of the clustering algorithm, redundant input parameters will lead to inaccurate results. Previous research has analyzed the influencing factors of the heterogeneity of the RMP and this population's housing demand. Through a survey of those relevant literatures, we obtained an ordered list according to the number of occurrences of the influencing factors (see Table 1). To improve the reasonability of the final empirical results, we filtered and modified 10 factors (Top 10) as research variables, based on previous findings and the purpose of this study, as shown in Table 1. In addition, the results of the basic descriptive statistical analysis are shown in Tables 2 and 3.

Table 1. Definitions of the influencing factors.

Influencing Factor	Parameter Type	Description	References
Age	continuous	2014–*birt_year* (years)	[36]
Education Status	categorical	The highest education that the interviewee has achieved includes; "primary school", "junior middle school", "senior high school", "associate's degree" and "bachelor's degree"	[37,38]

Table 1. *Cont.*

Influencing Factor	Parameter Type	Description	References
Marital Status	categorical	Includes "first marriage", "unmarried" and "others"	[34,39]
Mobile Duration	continuous	2014–*flo_year* (years)	[40]
Employment Identity	categorical	Occupation classification and role of the respondent in his work; includes "employee", "employer", "self-employed labourer" and "other"	[38]
Monthly Income	continuous	Average monthly gross income at the place of residence (RMB)	[34,38,40,41]
Housing Consumption	continuous	Average monthly housing rent at the place of residence (RMB)	[11,33,41]
Decision for Settlement	categorical	The answer to the question "Do you want to live locally for more than five years?"; includes "yes", "no" and "undefined"	[37,42]
Number of Minor Children	continuous	Number of children under 18 years of age raised by the respondent	[39,43]
Housing Nature	categorical	The nature of the respondent's dwelling; includes "rental house", "commercial house", "free house" and "other"	[37,44]

Notes: the *birt_year* represents the birth year of respondents; the *flo_year* indicates the migrating year of respondents. Source: Own elaboration.

Table 2. Descriptive statistics for continuous variables.

Variables	Mean	SD
Age	33.26	8.78
Mobile Duration	4.59	4.81
Monthly Income	2969.78	3658.91
Housing Consumption	334.83	446.00
Number of Minor Children	1.10	0.85

Source: Own elaboration.

Table 3. Descriptive statistics for categorical variables.

Variables	Frequency (%)	Variables	Frequency (%)
Education Status		**Employment Identity**	
Primary school	10.85	Employee	59.57
Junior middle school	55.01	Employer	8.50
Senior high school	22.94	Self-employed labourer	25.12
Associate's degree	8.50	Other	6.81
Bachelor's degree	2.69	**Housing Nature**	
Marital Status		Rental house	66.89
Unmarried	21.60	Commercial house	10.49
First marriage	75.96	Free house	12.46
Other	2.44	Other	10.16
Decision for Settlement			
Yes	57.45		
No or undefined	41.83		

Source: Own elaboration.

4. Method and Empirical Analysis

4.1. Partitioning around the Medoid (PAM)

The PAM algorithm, a variant of the K-medoids algorithm, is a clustering algorithm with the partitioning method. It is crucial for PAM to determine the center of each cluster. Compared with the

traditional K-means algorithm, the PAM algorithm chooses "medoids" instead of the "mean value" to represent the center of the cluster; thus, PAM is less sensitive to outliers and noise data [45], and has better clustering accuracy. In addition, the algorithm has the advantages of simple implementation, low space complexity, and stronger applicability to various types of data [46]. While there is a positive correlation between the time complexity of the algorithm and the sample size, the PAM algorithm is still suitable for solving cluster analysis problems of multivariate cross-sectional data [46]. In the first step of PAM, the initial medoids of each cluster are randomly generated based on the number of clusters (generally expressed in k). To reduce the errors of the objective function, the algorithm iteratively modifies the medoids. The final clustering results are obtained when the objective function is optimal.

In previous studies, this algorithm has been widely applied to different fields [47–49]. Different packages can be directly applied in many kinds of software for data analysis. Therefore, the detailed procedure will not be described here. For more information about the principle and implementation of PAM, please refer to a previous study [50].

4.2. Between-Class Distance Computation

In the process of clustering analysis, after choosing the clustering algorithm, we have to select the calculation method of the between-class distance according to the characteristics of our data. It is particularly important to choose the most suitable between-class distance computation, which can not only improve the credibility of the final clustering results, but also make the realistic conclusions drawn from the results more valuable. There are various kinds of between-class distance computations, for instance, Euclidean distance, Manhattan distance, and Chebyshev distance. Considering that there are two types of data (namely, numbers and characters) in the dataset of this paper, we attempted to overcome this problem by introducing Gower's dissimilarity (GD) coefficient into the empirical model.

GD, a method used to measure the similarity among different samples, was proposed by J.C. Gower in 1971. The model introducing GD offers us more choices for determining a cluster algorithm because it can measure the between-class distance of a dataset with both continuous and categorical data.

According to Gower [51], sample x_i and x_j are rewritten as $x_i = (x_{i1}, x_{i2},..., x_{in})'$ and $x_j = (x_{j1}, x_{j2},..., x_{jn})'$, respectively, and k is defined as the size of the dimension. For continuous data, the GD between sample i and sample j can be calculated by the following equation (1):

$$s_{ij} = \frac{1}{n}\sum_{k=1}^{n}\left(1 - \frac{\left|x_{ik} - x_{jk}\right|}{R_k}\right), \tag{1}$$

where R_k is the difference between the maximum and the minimum on the k dimension of sample x. For categorical data, variables with m categorical attributes are broken down into m 0–1 variables. Then, the distances between each sub-variable are weighted and summed by using the strategy of the Dice coefficient [52] to obtain the final GD between samples i and j, by which the following equation (2) can be calculated:

$$s_{ij} = \sum_{k=1}^{n} s_{ijk}\delta_{ijk} / \sum_{k=1}^{n} \delta_{ijk}, \tag{2}$$

where s_{ijk} and δ_{ijk} are the markup and weight values, respectively, whose values depend on the value of certain attributes. The markup and weight values of different cases are reported in Table 4.

Table 4. Markup and weight values.

Attributes of Sample *i*	Y	Y	F	F
Attributes of Sample *j*	Y	F	Y	F
s_{ijk}	1	0	0	0
δ_{ijk}	1	1	1	0

Source: Own elaboration based on [51].

4.3. Number of Clusters k

In a practical clustering analysis process, the *k* value, which is the number of clusters we initially set, has a profound impact on the final clustering result. It is rather difficult to determine the appropriate *k* value beforehand. To address this problem, a series of solutions based on practical experience and speculative knowledge have been proposed. On the one hand, the *k* value must be less than the number of characteristic variables, because it is hoped that features of each cluster are part of the characteristic variables of the source data. On the other hand, a number of statistical indices, such as the Silhouette coefficient (SC) index, have been proposed to estimate the clustering performance with different *k* values [53].

Based on these solutions, we set a possible range of *k* values, according to previous studies and the idiographic conditions of the dataset, after which the clustering results are evaluated by the SC. The calculations of the SC often involve the individual SC and the global SC. For each sample in the dataset, the individual SC can be expressed as:

$$s_i = \frac{b_i - a_i}{\max(a_i, b_i)}, \tag{3}$$

where the subscripts *i* indicate the sample, a_i is the average distance between sample *i* and other samples belonging to the same cluster as sample *i*, and bi is the average distance between sample *i* and the sample belonging to the most similar but different clusters of sample *i*. According to Equation (3), the maximum and minimum values of s_i are 1 and -1, respectively, and the closer s_i is to 1, the more accurate the clustering result is. The global SC is the average of the individual SC of all samples. We can use it to evaluate the rationality of the clustering results. The larger the value is, the better the performance.

4.4. A Mixed-Type Data Clustering Analysis Model

The general solution to mixed type data clustering analysis consists of two steps: A similarity measure, and algorithm selection. For the purposes of this study, an improved model based on PAM for solving the mixed type data clustering problem is provided. In this model, the similarity of samples is calculated by using the Gower coefficient. Then, the clustering experiments are carried out with different initial numbers of clusters, after which the clustering results are verified based on the theory of the SC. The best result has the characteristics of the maximum global SC value. The flowchart of the model is shown in Figure 1.

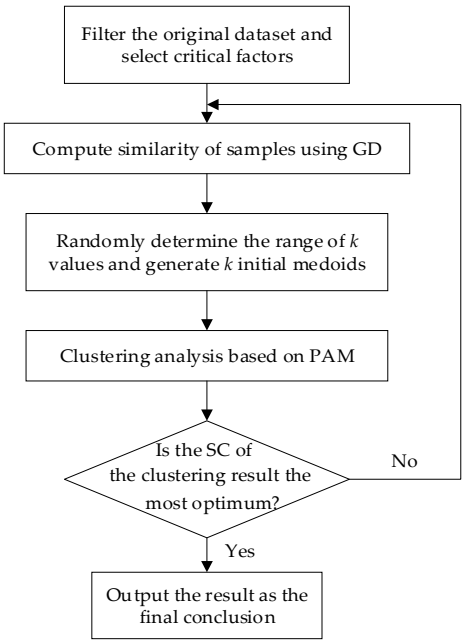

Figure 1. Flowchart of the model. Source: Own elaboration.

4.5. Empirical Analysis

Both practical experience and data regularity should be considered when the *k* value is determined subjectively. Previous researchers have suggested that two to five clusters might realistically represent the RMP class differentiation [54]. Based on the requirements of the model that we discussed in Chapter 4, the initial k value must be less than the number of influencing factors that we selected. When the *k* value varies in the interval [2,10], the dataset is analyzed by using the model that we proposed. The SC values of each clustering result are calculated and compared, and the results are displayed with the curve in Figure 2.

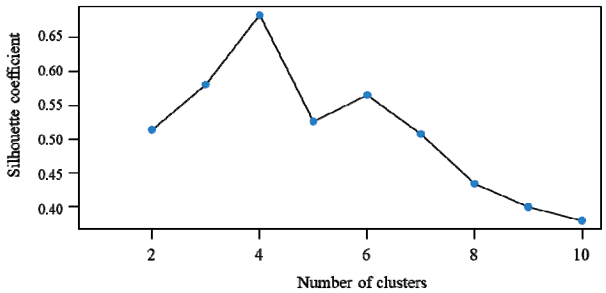

Figure 2. Line chart of the Silhouette coefficient with different numbers of clusters. Source: Own elaboration.

The best k value can be obtained from Figure 2. When $k < 4$, the SC is positively correlated with the *k* value, which indicates that the clustering result is closer to reality. Nevertheless, the SC falls sharply when $k = 5$. After SC increases, it falls after $k > 6$. According to the above analysis, we find that the SC peaks at 0.683 when $k = 4$, which means that it is most ideal to aggregate the overall samples into four clusters.

RStudio v.1.1.456 (Copyright RStudio Inc., Boston, MA, USA), an integrated development environment for R language, was used for statistical analysis. The terse and effective programming environment with code editing is established by providing extensive program packages to reduce the time cost of modelling. There are two functions in the cluster package: *daisy* and *pam*. The former can

be used to compute the pairwise dissimilarities between samples with mixed type variables, while the latter is a function implementation of the PAM algorithm. Furthermore, the SC can be acquired from the results computed by the pam function.

5. Empirical Results and Discussion

5.1. Population Clustering Results based on Housing Demand

Empirical analysis of the 10 variables of the RMP in our study is presented in the descriptive statistical summary in Table 5. To highlight the unique characteristics of each cluster, and to make the results more intuitive, some variable groupings with an extremely small sample sizes are merged or modified.

Table 5. Heterogeneous structural features of four rural migrant population (RMP) clusters (%).

Variable		#1	#2	#3	#4
Age	15–24	62.00	1.75	1.71	4.75
	25–34	27.74	19.79	6.80	9.83
	35–44	8.66	50.12	29.33	18.73
	45–54	1.07	27.45	50.82	52.27
	More than 55	0.53	0.89	11.34	14.42
Education Status	Primary school	2.17	0.16	80.33	1.01
	Junior middle school	86.28	2.09	15.29	12.26
	Senior high school	8.95	71.73	3.01	76.14
	Associate degree	2.59	11.51	1.37	5.84
	Bachelor's degree	0.01	14.51	0.00	4.75
Marital Status	Unmarried	75.03	9.83	3.87	3.74
	First marriage	22.72	87.67	91.07	86.62
	Other	2.25	2.50	5.06	9.64
Mobile Duration	Low level	76.35	9.11	82.46	10.96
	Average level	14.01	83.27	15.94	14.21
	High level	9.64	7.62	1.60	74.83
Employment Identity	Employee	88.45	66.69	81.84	0.00
	Employer	5.25	15.07	7.39	14.00
	Self-employed labourer	0.00	8.49	1.00	79.11
	Other	6.30	9.75	9.77	6.89
Monthly Income	Low level	65.26	14.63	74.97	5.27
	Average level	25.35	57.46	19.49	29.14
	High level	9.39	27.91	5.54	65.59
Housing Consumption	Low level	77.47	51.71	71.95	48.68
	Average level	14.11	26.84	11.74	22.63
	High level	8.42	21.45	16.31	28.69
Decision for Settlement	Yes	11.48	70.04	0.81	86.18
	No or undefined	88.52	29.96	99.19	13.82
Number of Minor Children	0	72.14	18.60	40.42	56.87
	1 or 2	26.05	77.38	36.16	32.58
	More than 2	1.81	4.02	23.42	10.55
Housing Nature	Rental house	62.36	78.13	66.23	62.07
	Commercial house	8.98	13.31	8.83	12.80
	Free house	19.11	1.25	15.05	16.69
	Other	9.55	7.31	9.89	8.44

Notes: #1~#4 represents the first to the fourth population after clustering analysis respectively. Source: Own elaboration.

As shown in Table 5, it may be inferred that the class stratification of the RMP is mainly manifested in the following aspects.

First, there are some common characteristics among the four clusters. Populations with low housing consumption levels and personal housing renters make up a relatively high proportion in all four clusters of the RMP. This indicates that the living conditions of the RMP are likely to be poor, as most of this population has little capacity to pay for houses in the city. Worse, this difficult situation has not improved.

Second, while the above characteristic is common to the four groups, other characteristics display significant differences. For instance, the population of the first cluster (#1) is similar to that of the third cluster (#3) in terms of income level, education level, employment status, and settlement intention. The difference between them is that the former consists mainly of the new generation, and the latter is elderly. Table 5 shows the descriptions of the four clusters. Next, some preliminary conclusions are presented. The population of the first cluster, with an average age of 26.21 years, has a considerable proportion with a junior high school education, is generally engaged in service occupations, and has a low income level, on average. However, due to the small average household dependency burden, the level of household consumption is lower. Characteristics of the individual and of the urban migration are deficient in this group. The third cluster, which is limited by a low education and an advanced age of 47.90 years, on average, obtains a meagre remuneration by participating in manual labor (such as construction work). Because of weighty family obligations, the people in this cluster are forced to hunt for a better-paying job in the city. When their living conditions improve, they will most likely return to the countryside to settle down, because they lack identity and belonging in the urban setting. According to the above analysis, the first and third clusters are defined as "new generations just entering" (#1) and "old generations just entering" (#3), respectively.

Third, there are some symbolic characteristics of each cluster. Taking the second cluster (#2) and the fourth cluster (#4) as examples, the keywords of the second cluster include middle-aged, married (at least 87.67%) with a child (81.4%), a talent in a technical profession (analysis based on the original dataset), and a higher level of education (71.73% with senior high school education). Therefore, the proportion of this cluster that is engaged in a technology-based occupation is larger, and the income is higher. Further analysis based on the original dataset indicates that the migration pattern of this cluster is family migration, involving peasant workers who migrate with their children. This group pays more attention to the supporting public resources than to other aspects, and has a greater possibility of settling in urban cities. The age composition of the fourth cluster is more complex and consists chiefly of middle-aged and elderly people. Most of them work as employers and self-employed workers, and they have a strong desire for settlement in urban cities. Either the income level or the social position of the population of the fourth cluster is higher than that of the people in the other clusters. Their long-term work experience and strong city identity make it possible for these people to hold a stable job and to live comfortably in urban cities. These are the reasons for why their homeownership rate is the highest among the four clusters. According to the above analysis, the second and fourth clusters are defined as being "technical employee class" (#2) and "employer or self-employed class" (#4), respectively.

5.2. Distribution of RMP Groups in Various Cities

In the process of the formulation and implementation of social security policies in China, obvious regional distribution differences are found between various clusters; that is, the population distribution of these clusters differs across cities [55]. Practical experience suggests that the social security policy in China should be specifically devolved to the local level to balance the urban sustainability [56,57]. At present, there are obvious weaknesses in housing security for RMP in China. Both urban residents and RMP are regarded as security objects, based on the interim measures for the management of public rental housing, published in 2012. However, the major housing security resources are applicable for the urban population because of their urban identity (known in China as "urban hukou"). In fact, the problems of housing security policy for RMP, especially overlooking the regional features of RMP between various cities, still exist. It is essential to observe the regional distribution of the RMP from the city perspective as

an effective measure to increase housing policy efficiency at the stage of macroeconomic policy-making. The State Council measures the urban population scale, based on the resident population size, to classify cities as super cities, mega cities, large cities, middle cities, and small cities [58]. In this study, the numbers of primary sample units per city classification are 5, 8, 14, 40, and 15, respectively. The spatial distribution of the population at the city level is reported in Table 6.

Table 6. The proportion of the four RMP clusters in various cities (%).

	Super Cities	Mega Cities	Large Cities	Middle Cities	Small Cities	Total
#1	37.59	34.43	17.73	6.39	3.86	100.00
#2	10.91	41.56	9.60	35.39	2.54	100.00
#3	28.17	38.61	22.31	8.34	2.57	100.00
#4	27.69	25.45	20.60	16.36	9.90	100.00
Total	26.09	35.01	17.56	16.62	4.72	100.00

Note: The result is the structure of the RMP in different urban types, based on Table 5. The meaning of each cluster (#1, #2, #3, #4) is the same as in Section 5.1. Source: Own elaboration.

From the city perspective, it is obvious that major cities (namely, super cities, mega cities, and large cities), and especially super cities, attract the main portion of the RMP. On the one hand, many opportunities for employment are provided in major cities, which creates considerable private wealth to meet the household needs of low-income RMP families. On the other hand, there are many more housing security resources and public service resources in major cities than in other cities, which are more attractive for the long-term development of the RMP. However, blind population mobility may be responsible for aggravating the contradiction between rapid population influx growth and housing shortage, which is detrimental to the coordinated and orderly development of cities. Consequently, the countermeasures for major cities and medium-sized cities should be different. For major cities with strong population pressure, the inflow of the RMP should be restricted reasonably, and the threshold of housing security should be raised appropriately. However, positive housing security policies are appropriate for small and medium-sized cities with vast development potential, which contributes to enhancing their attractiveness. In other words, housing security policy should be initiated, based on the local city situation, to guide the reasonable flow of the RMP, which may effectively relieve pressure on population and housing in major cities. At the same time, this would promote the sustainable development of small and medium-sized cities.

From the perspective of the population, the characteristics of settlements in large cities is demonstrated by the first and the third clusters, which are defined by their original intentions to migrate. People in these clusters migrate to cities in search of more job opportunities and higher income levels. To accumulate wealth rapidly, these people may be more willing to compromise housing conditions. Makeshift houses, villages in cities, and shanty towns are currently the primary housing conditions of the people in these clusters. An effective approach to improving the situation is by providing more policy-based public housing with strict management roles. In contrast, most of the people in the second cluster are married, and favor mega-cities and middle cities as migrant destinations. Both harmonious family life and high income are highly valued by those in this cluster, which makes services such as educational resources and employment opportunities the key factor in their housing choice. The reality in China, however, is that major cities have more social public security resources, but the access threshold for housing buyers is relatively low in small and medium-sized cities. As a result, mega-cities become the best place for these people to migrate, while people who are limited by insufficient housing affordability will migrate to middle cities as an alternative. It is critical to distribute social public security resources fairly, and to provide better economic conditions and policy support to small- and medium-sized cities, which can promote the balanced development of urbanization between cities of different sizes. Moreover, this is an essential requirement for the sustainability of urban housing.

6. Conclusions and Suggestions

In this paper, a model for mixed-type data clustering analysis combined with the GD coefficient and PAM algorithm was proposed. Empirical analysis was conducted to explore the demographics of the RMP from a sustainable urban development perspective, based on the microdata of the RMP in eastern China. In the research results, four clusters of the RMP were identified: "new generations just entering", "old generations just entering", "technical employee class", and "employer or self-employed class". In general, the consumption level of all four clusters is low, and they prefer rental housing to other forms of housing, partly due to their meagre income. Further analysis indicates that large-sized cities are much more attractive to the RMP, while different clusters have unique preferences concerning the place to settle.

Some suggestions for the formulation of housing security policy based on this study are presented, to promote sustainable urban development.

At present, the contradictions between the single supply mode of housing security and the varied housing demand of the RMP have become the main problem in the field of the sustainable development of urban housing, which requires a diversified housing security system that is based on population stratification. Previous researchers tend to regard the RMP as a homogeneous group when making recommendations for building the housing security system. With the advancement of urbanization, demographic stratification has emerged within the RMP. The diversity of actual housing demand caused by social status and household background is reflected in intrinsic and extrinsic factors, which include age, education, occupation, duration of mobility, city identity, and income. The unitary housing security pattern cannot solve the current challenges, which may be resolved by diverse modalities, particularly the preferential policy of public housing. The most fundamental measure for the relevant administrative departments is to analyze and explore the actual housing security demand of the RMP, based on diverse group features, which is conducive to improving the housing affordability of the RMP through multiple approaches.

From the perspective of the city, both the demographic structure and the local regional features should be considered comprehensively in the period of designing housing security policy. Meanwhile, it is necessary to ensure that the common and individual demands are embodied in the policy, which can play an active role in sustainable urban development and urbanization. Metropolises, such as Beijing, have many employment opportunities and income levels, which have a strong appeal for the RMP who are interested in high income but who lack city identity. However, a consequence of this is that the serious contradictions in housing have caused a decline in living standards. Tighter policies for housing security can be adopted to limit the population inflow. We suggest that the access threshold of policy could be raised appropriately. The applicant satisfies the criteria of local urban residents who join social insurance schemes and sign work contracts. In contrast, the housing resources of small- and medium-sized cities cannot be fully utilized, because of the insufficient urban attractiveness of these cities. We suggest that the application conditions of security housing should be diluted. Meanwhile, we can explore diverse forms of housing security, such as housing subsidies and monetary subsidies. Increasing the supply of public housing can make it possible to incorporate RMP into the urban housing security system. Therefore, effective policy for housing security consists of multiple ways to meet diverse needs and to promote sustainable city development.

The group of participants that is concerned with the housing security policy is large and complex. Because such policy concerns distribution, there may be information gaps. Faced with this situation, the dominant approaches are to establish methodical cooperation between suppliers of RMP housing security, and to strengthen the interdepartmental sharing of information resources, which is beneficial for the improvement of governance capacity. Therefore, the irregular changes in the housing demand of the RMP should be analyzed based on the population's demographic structure. In addition, the important prerequisite for solving the series of social security problems for RMP, especially the problem of sustainable urban housing, is to build a big data management platform that is based on information related to housing security and RMP households. On the one hand, it is useful to

achieve the pluralistic management of housing security resources and the effective use of public service resources. On the other hand, it is beneficial to make the housing security policies more effective and the resource distribution more rational.

Author Contributions: The paper was conducted by X.D., L.G., Y.G., X.C. and K.X.; X.D. conceived and design the study. L.G. analyzed the data and wrote the paper. Y.G. debugged the computer codes for the algorithm. X.C. and K.X. critically revised it for important content.

Funding: This study was funded by the Humanities and Social Science project of the Chinese Ministry of Education [Grant No. 14YJA630006], Key Research and Development Programme of Shandong Province [Grant No. 2018GGX106006] and the University in Shandong Province of Humanities and Social Science project [Grant No. J16YF14].

Acknowledgments: The authors would like to thank the experts who reviewed the manuscript.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Qu, Z.F.; Zhao, Z. Evolution of the Chinese rural-urban migrant labor market from 2002 to 2007. *China Agric. Econ. Rev.* **2014**, *6*, 316–334.
2. Ngai, P.; Lu, H.L. Unfinished Proletarianization: Self, Anger, and Class Action among the Second Generation of Peasant-Workers in Present-Day China. *Mod. China* **2010**, *36*, 493–519. [CrossRef]
3. Mok, K.H.; Ngok, K.J.W. A new working class in the making? The rise of the peasant workers and implications for social policy in China. *China Q.* **2011**, *38*, 241.
4. Wang, C.C.; Jing, C. Contribution to Economic Growth from Rural-Urban Migrant Workers and the Income Share in the Process of Urbanization: Evidence from 1995–2009 in China. *Actual Probl. Econ.* **2012**, *135*, 245–254.
5. Annual Report of Migrant Workers Survey in 2017. Available online: http://www.stats.gov.cn/tjsj/zxfb/201804/t20180427_1596389.html (accessed on 25 August 2018).
6. Niu, G.; Zhao, G.C. Living condition among China's rural-urban migrants: Recent dynamics and the inland-coastal differential. *Hous. Stud.* **2018**, *33*, 476–493. [CrossRef]
7. Chu, R.; Liu, M.; Shi, G.J. How rural-urban identification influences consumption patterns? Evidence from Chinese migrant workers. *Asia Pac. J. Mark. Logist.* **2015**, *27*, 40–60. [CrossRef]
8. Wu, W.P.; Wang, G.X. Together but Unequal: Citizenship Rights for Migrants and Locals in Urban China. *Urban Aff. Rev.* **2014**, *50*, 781–805. [CrossRef]
9. Zong, X.L.; Guan, X.; Gao, Y.; Chen, Z.L.; Zhang, G.X. Analysis of the Influencing Factors of Migrant Worker Social Insurance in Lanzhou. *Eurasia J. Math. Sci. Technol. Educ.* **2017**, *13*, 7949–7960. [CrossRef]
10. Niu, Y. The performance and problems of affordable housing policy in China: The estimations of benefits, costs and affordability. *Int. J. Hous. Mark. Anal.* **2008**, *1*, 125–146. [CrossRef]
11. Tian, X.; Hu, J.; Zhang, C.; Zhao, Y. Housing expenditure and home purchasing behaviors of rural-urban migrants in China. *China Agric. Econ. Rev.* **2017**, *9*, 558–566. [CrossRef]
12. Jia, M.; Heath, T. China's diversifying demand for housing for the elderly. *Int. J. Hous. Mark. Anal.* **2016**, *9*, 256–271. [CrossRef]
13. Poon, J.; Garratt, D. Evaluating UK housing policies to tackle housing affordability. *Int. J. Hous. Mark. Anal.* **2012**, *5*, 253–271. [CrossRef]
14. Xing, C. Migration, self-selection and income distributions. *Econ. Transit.* **2014**, *22*, 539–576. [CrossRef]
15. Snel, E.; Faber, M.; Engbersen, G. Civic Stratification and Social Positioning: CEE Labour Migrants without a Work Permit. *Popul. Space Place* **2015**, *21*, 518–534. [CrossRef]
16. van Leeuwen, M.H.D.; Maas, I. Historical Studies of Social Mobility and Stratification. *Annu. Rev. Sociol.* **2010**, *36*, 429–451. [CrossRef]
17. Jianshe, X. Stratification of migrant rural workers in the urbanization of China. *J.Guangzhou Univ.(Soc. Sci. Ed.)* **2006**, *10*, 44–49. (In Chinese)
18. He, S.J.; Liu, Y.T.; Wu, F.L.; Webster, C. Social Groups and Housing Differentiation in China's Urban Villages: An Institutional Interpretation. *Hous. Stud.* **2010**, *25*, 671–691. [CrossRef]
19. Kepinska, E. Gender Differentiation in Seasonal Migration: The Case of Poland. *J. Ethn. Migr. Stud.* **2013**, *39*, 535–555. [CrossRef]

20. Mora, C.; Undurraga, E.A. Racialisation of Immigrants at Work: Labour Mobility and Segmentation of Peruvian Migrants in Chile. *Bull. Lat. Am. Res.* **2013**, *32*, 294–310. [CrossRef]
21. Greenwood, M.J. Modeling the age and age composition of late 19th century US immigrants from Europe. *Explor. Econ. Hist.* **2007**, *44*, 255–269. [CrossRef]
22. De Alwis, S.; Parr, N. Differences in occupation between ancestry subgroups of Asian birthplace groups in Australia. *Aust. J. Soc. Issues* **2018**, *53*, 173–202. [CrossRef]
23. Wang, D.H.; Yang, X.J.; Hao, F.J. A study on the spatial characteristics and correlation of migrant workers' urban integration and well-being: A case study of Xi'an (China). In Proceedings of the 3rd International Conference on Agricultural and Biological Sciences, Qingdao, China, 26–29 June 2017; Iop Publishing Ltd.: Bristol, UK, 2017; Volume 77.
24. Winston, N. Regeneration for Sustainable Communities? Barriers to Implementing Sustainable Housing in Urban Areas. *Sustain. Dev.* **2010**, *18*, 319–330. [CrossRef]
25. Zhu, Y. China's floating population and their settlement intention in the cities: Beyond the Hukou reform. *Habitat Int.* **2007**, *31*, 65–76. [CrossRef]
26. Kagochi, J.M.; Mace, L.M. The determinants of demand for single family housing in Alabama urbanized areas. *Int. J. Hous. Mark. Anal.* **2009**, *2*, 132–144. [CrossRef]
27. Öztürk, A.; Kapusuz, Y.E.; Tanrıvermiş, H. The dynamics of housing affordability and housing demand analysis in Ankara. *Int. J. Hous. Mark. Anal.* **2018**, *11*, 828–851. [CrossRef]
28. Apergis, N.; Rezitis, A. Housing prices and macroeconomic factors in Greece: Prospects within the EMU. *Appl. Econ. Lett.* **2003**, *10*, 561–565. [CrossRef]
29. Mankiw, N.G.; Weil, D.P. The Baby Boom, The Baby Bust, and the Housing Market. *Natl. Bur. Econ. Res.* **1989**, *19*, 235–258. [CrossRef]
30. Han, X.H. Housing demand in Shanghai: A discrete choice approach. *China Econ. Rev.* **2010**, *21*, 355–376. [CrossRef]
31. Eichholtz, P.; Lindenthal, T. Demographics, human capital, and the demand for housing. *J. Hous. Econ.* **2014**, *26*, 19–32. [CrossRef]
32. Flambard, V. Demand for housing choices in the north of France: A discrete approach. *J. Eur. Real Estate Res.* **2017**, *10*, 346–365. [CrossRef]
33. Lu, P.; Zhou, T. Housing for Rural Migrant Workers: Consumption Characteristics and Supply Policy. *Urban Policy Res.* **2008**, *26*, 297–308. [CrossRef]
34. Oktay, E.; Karaaslan, A.; Alkan, Ö.; Kemal Çelik, A. Determinants of housing demand in the Erzurum province, Turkey. *Int. J. Hous. Mark. Anal.* **2014**, *7*, 586–602. [CrossRef]
35. China Migrants Dynamic Survey. Available online: http://www.chinaldrk.org.cn/wjw/#/data/classify/population/yearList. (accessed on 19 June 2018).
36. Chen, S.Q. Study on intergenerational differences between the new generation of migrant workers and the first generation of migrant workers. In Proceedings of the 2014 International Conference on Education, Management and Computing Technology, Xiamen, China, 22–23 November 2014; Zhang, H.M., Ed.; Atlantis Press: Paris, France, 2014; Volume 100, pp. 468–472.
37. Yujun, W.; Wenhui, Y.; Zhilin, L. Determinants of Changes in Housing Sources for Migrant Workers in Urban China: An Empirical Study Based on a Twelve-city Migrant Survey. *Popul. Res.* **2014**, *38*, 63–74. (In Chinese)
38. Elisabeth Birkelund, G.; Lemel, Y. Lifestyles and social stratification: An explorative study of France and Norway. In *Class and Stratification Analysis*; Emerald Group Publishing Limited: Bingley, UK, 2013; pp. 189–220.
39. Xiang, H.L.; Yang, J.; Zhang, T.P.M.; Ye, X.Y. Analyzing In-Migrants and Out-Migrants in Urban China. *Appl. Spat. Anal. Policy* **2018**, *11*, 81–102. [CrossRef]
40. Zhuang, M. The Social Support Network for Rural Migrant Workers in Chengdu, China: Local Governance and Civil Society in the Fight Against Poverty and Exclusion. *Ids Bull. Inst. Dev. Stud.* **2009**, *40*, 41–49. [CrossRef]
41. Rabe, B.; Taylor, M.P. Differences in Opportunities? Wage, Employment and House-Price Effects on Migration. *Oxf. Bull. Econ. Stat.* **2012**, *74*, 831–855. [CrossRef]
42. Xie, S.H.; Chen, J. Beyond homeownership: Housing conditions, housing support and rural migrant urban settlement intentions in China. *Cities* **2018**, *78*, 76–86. [CrossRef]

43. Liu, Y.; Li, Z.C. Determinants of Housing Purchase Decision: An Empirical Study of the High Education Cohort in Urban China. *J. Asian Archit. Build. Eng.* **2018**, *17*, 299–305. [CrossRef]
44. Fougere, D.; Kramarz, F.; Rathelot, R.; Safi, M. Social housing and location choices of immigrants in France. *Int. J. Manpow.* **2013**, *34*, 56–69. [CrossRef]
45. Arora, P.; Deepali, V.; Varshney, S. Analysis of K-Means and K-Medoids Algorithm For Big Data. In Proceedings of the 1st International Conference on Information Security & Privacy, Angers, France, 9–11 February 2015; Abraham, J., Bhatnagar, V., Eds.; Elsevier Science B.V.: Amsterdam, The Netherlands, 2016; Volume 78, pp. 507–512.
46. Li, Z.M.; Wang, G.F.; He, G.Y. Milling tool wear state recognition based on partitioning around medoids (PAM) clustering. *Int. J. Adv. Manuf. Technol.* **2017**, *88*, 1203–1213. [CrossRef]
47. Ignaccolo, R.; Ghigo, S.; Giovenali, E. Analysis of air quality monitoring networks by functional clustering. *Environmetrics* **2008**, *19*, 672–686. [CrossRef]
48. Hernandez-Torruco, J.; Canul-Reich, J.; Frausto-Solis, J.; Mendez-Castillo, J.J. Feature Selection for Better Identification of Subtypes of Guillain-Barre Syndrome. *Comput. Math. Methods Med.* **2014**, *9*. [CrossRef] [PubMed]
49. Wang, K.J.; Zheng, J.; Zhang, J.Y.; Dong, J.Y. Estimating the Number of Clusters via System Evolution for Cluster Analysis of Gene Expression Data. *IEEE Trans. Inf. Technol. Biomed.* **2009**, *13*, 848–853. [CrossRef] [PubMed]
50. Kaufman, L.; Rousseeuw, P.J. *Finding Groups in Data: An Introduction to Cluster Analysis*; John Wiley & Sons: Hoboken, NJ, USA, 2009; Volume 344.
51. Gower, J.C. A general coefficient of similarity and some of its properties. *Biometrics* **1971**, 857–871. [CrossRef]
52. Dice, L.R. Measures of the amount of ecologic association between species. *Ecology* **1945**, *26*, 297–302. [CrossRef]
53. Ayton, R.L.; Watters, P.; Dazeley, R. Evaluating authorship distance methods using the positive Silhouette coefficient. *Nat. Lang. Eng.* **2013**, *19*, 517–535.
54. Jianping, L.; Pengpeng, Y. An Analysis Framework and Empirical of Internal Differentiation of Migrant Workers. *Reform Econ. Syst.* **2015**, *5*, 98–104. (In Chinese)
55. Hao, P.; Tang, S.S. Migration destinations in the urban hierarchy in China: Evidence from Jiangsu. *Popul. Space Place* **2018**, *24*, 14. [CrossRef]
56. Li, Q. Research on Implementation of Frontline Public Policy in Current China. In Proceedings of the 2nd International Conference on Contemporary Education, Social Sciences and Humanities, Moscow, Russia, 14–15 June 2017; McAnally, E., Zhang, Y., Green, R., Tretyakova, I., Eds.; Atlantis Press: Paris, France, 2017; Volume 124, pp. 946–949.
57. Andersson, K.; Kalman, H. Methodological challenges in the implementation and evaluation of social welfare policies. *Int. J. Soc. Res. Methodol.* **2012**, *15*, 69–80. [CrossRef]
58. Notice on Adjusting the Scale of Urban Scale. Available online: http://www.gov.cn/zhengce/content/2014-11/20/content_9225.htm (accessed on 3 September 2018).

© 2018 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (http://creativecommons.org/licenses/by/4.0/).