

Article

Predicting Convergence of Per Capita Income in Spain: A Markov and Cluster Approach

José F. Gálvez-Rodríguez * , Miguel Manzano-Hidalgo and Amelia V. García-Luengo * 

Department of Mathematics, University of Almería, 04120 Almería, Spain; miguelmh97@gmail.com

* Correspondence: jgr409@ual.es (J.F.G.-R.); amgarcia@ual.es (A.V.G.-L.)

Abstract: In this work we analyze the evolution of productivity, in terms of the convergence of per capita income, of all the Spanish provinces, based on data from the previous decade. On the one hand, a cluster analysis allows us to group the Spanish provinces according to four income levels (low, medium-low, medium-high and high), which can be determined from the quartiles of the distribution, and, on the other hand, Markov chains make it possible to study the long-term evolution of productivity and convergence between the provinces, as well as the speed of convergence towards the equilibrium situation. Moreover, we can obtain the average time to return to an income level in which a province was previously. With the above, predictions of future income levels are made for the provinces, both in the current situation, and if the pandemic caused by COVID-19 had not existed, which leads us to evaluate the impact of the health emergency.

Keywords: cluster analysis; Markov chain; per capita income



Academic Editor: Fabio Clementi

Received: 2 December 2024

Revised: 7 January 2025

Accepted: 8 January 2025

Published: 11 January 2025

Citation: Gálvez-Rodríguez, J.F., Manzano-Hidalgo, M., García-Luengo, A.V. (2025). Predicting Convergence of Per Capita Income in Spain: A Markov and Cluster Approach. *Economics*, 13(1), 17. <https://doi.org/10.3390/economics13010017>

Copyright: © 2025 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 (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Markov chains Takács (1960) are an important mathematical tool in the analysis of time series and have been used by several authors when studying the economic growth and other phenomena, even in other disciplines. Markov analysis starts from the present value of a certain variable in order to predict its future. For example, Quah (1993) uses Markov chains to analyze how countries exchange positions over time in terms of their GDP per capita. Each finite and homogeneous Markov chain is characterized by a matrix which contains different probabilities. In the study cited above, the probabilities that a country remains in its current income group or moves to a higher or lower income group are calculated and collected in this matrix. It is concluded that, in general, countries tend to converge in terms of their GDP per capita over time, but the rate of convergence varies depending on the region and the time period studied. This approach has been used by other authors in similar analyses in a wide variety of contexts. Some relevant examples of studies in which the Markov chain methodology has been used to analyze the dynamics of growth and convergence are the following:

- Fingleton (1997) uses the evidence given by data from 1975 to 1993 in order to justify that different regions of the European Union seem to be converging towards stable proportions in terms of per capita income levels.
- Bode and Nunnenkamp (2011) analyze the impact of foreign direct investment on per capita income and growth, in general, in the United States, since the mid-1970s, demonstrating that the investment in employment has favored the increase in income, although the investment destined for capital has not had the same impact on the states with the greatest poverty, where it has been most intensive.

- [Lipták \(2011\)](#) examines the evolution of unemployment in the Hungarian labor market during the period 1992–2009.

Despite being a methodology exposed at the end of the previous century, Markov chains keep on appearing in the nowadays literature to study the evolution of economic variables. We refer the reader to the works by [Wenxuan \(2023\)](#), [Rey \(2023\)](#), [Papanikolaou \(2020\)](#), [Arreola and Montiel \(2024\)](#), [Kerkouch et al. \(2024\)](#), [Chen et al. \(2022\)](#), [Karahasan \(2020\)](#) and [Haller et al. \(2020\)](#).

However, knowing the probability of transition between states along a period of time is not enough, since we still do not know which region continues in the same situation or change it after several periods of time. That is the reason why considering a cluster analysis can be of some interest in order to know, also, the elements of each group with respect to the past data. Cluster analysis, for what a good reference is [Everitt et al. \(2011\)](#), is a statistical technique whose primary objective is to group data according to their similarities and differences based on certain characteristics. It is a widely used method in Economics and Finance, to analyze the structure of markets, identify patterns and trends, and classify countries or companies into different groups in accordance with their economic and financial properties. A particular example of a study with cluster analysis is the one carried out by [Yang and Hu \(2008\)](#), who analyze data from the China Human Development Index in 1982, 1995, 1999 and 2003, in order to classify its provinces into four levels of income based on the three basic aspects contemplated in that index. We can also refer to other recent studies which show the interest of clustering in analyzing some economic variables. For example, [Gostkowski et al. \(2021\)](#) study the relationship between the national level of economic development and energy consumption in the main sectors of industry of the countries that belong to the Visegrad Group; [He et al. \(2021\)](#) analyze the socio-economic spatial structure of urban agglomeration in China by using clustering, and [Zarikas et al. \(2020\)](#) use clustering to group countries with respect to active cases, active cases per population and active cases per population and per area, so that the impact of the COVID-19 pandemic can be explored. In this sense, different economic variables such as GDP per capita, economic growth rate, foreign investment, international trade, among others, can be used to group countries into different categories based on their level of economic development and its growth dynamics. For example, cluster analysis can be used to identify groups of countries that converge with each other, that is, which have similar levels of economic development and are growing at a similar rate. In this way, as we will develop later, patterns of convergence or divergence between countries can be identified and the underlying causes of these trends can be analyzed.

We can conclude from this that cluster analysis is a useful tool for analyzing the structure and dynamics of countries, and can be used in combination with other statistical techniques, such as time series analysis and Markov chains, to obtain a more complete and accurate view of economic convergence between countries.

The main goal of this work is the application of the previously mentioned methodologies to the study of the convergence of per capita income of the Spanish provinces (in fact, we consider the fifty provinces together with the two autonomous cities in Spain) based on the GDP per capita of each of them in the years of the last decade. Moreover, we can evaluate the impact of COVID-19 pandemic on this convergence and study its speed of this convergence, both with and without COVID-19. Indeed, there are some recent works in which the authors face COVID-19 to income inequality, such as the one carried out by [Deaton \(2021\)](#). Additionally, a cluster analysis can let us have an idea of the possible groups of income in the future. Hence, the structure of the work is as follows: Section 2 provides some concepts and results that the reader should recall, which have to do with these two statistical methodologies as well as some reasons to chose both of them in the study. Moreover, Section 3 includes the application of them to the study of the evolution of

per capita income in Spain, by provinces, based on historical data, analyzing the behavior of the Spanish economy in the long term, both in the current situation and in case the pandemic generated by COVID-19 had not existed. Furthermore, in this section, a discussion is given together with the results. Finally, Section 4 collects the main conclusions of the work.

Provincial Income Disparities in Spain: A Literature Overview

As stated at the beginning of this section, Markov chains have become an interesting approach to deal with the evolution of economic variables and, in particular, of per capita income over the years, thanks to many works in the economic literature. In this subsection, we focus on Spanish studies which are related to this topic. On the one hand, it is worth mentioning the work by [Le Gallo and Chasco \(2008\)](#), who study the evolution of the population growth among the group of 722 municipalities included in the Spanish urban areas over the period 1900–2001. Furthermore, [Ayuda et al. \(2010\)](#) analyze the disparities in long-run regional population growth in continental Europe, concluding that there is a common pattern of divergence in economic growth of Europe. Markov chains are used in this paper and they also consider Spain as a particular case in the study, for what a cluster analysis is useful. On the other hand, [Gardeazábal \(1996\)](#) studies the Spanish provinces dynamics, in terms of their income, in the time period between 1967 and 1991, concluding that they tend to the equilibrium distribution, being more concentrated in medium levels of income. However, [Tirado et al. \(2016\)](#) explore per-capita GDP disparities across Spanish provinces from 1860 to 2010. The previous cited works suggest us considering Markov chains and cluster analysis in order to study the evolution of per capita income in Spain for the last decade, which has not been analyzed in the literature yet.

2. Methodology

In this section, we collect some concepts and results related to Markov chains and cluster analysis, which will be used in the following section to get the results of the main study. Moreover, we introduce the data we will work with and discuss the hypothesis on the Markov chains approach and the suitability of the type of clustering used in the study.

2.1. Markov Chains

We refer the reader to Appendix A in order to recall some basic preliminaries on Markov chains. In this paper, we will work with finite and homogeneous discrete-time Markov chains. The Markov chain must be finite because we want to group the provinces into a finite number of income states, as usual in the related literature. Finally, homogeneity assumption has to do with the idea of getting a unique matrix which gathers all the information according to the collected data.

2.1.1. Long-Term Behavior

Dynamic models which are based on Markov chains have, as a point of interest, the analysis of the convergence of the transition probabilities when the time tends to infinity. This leads us to study the long-term behavior, also known as stationary or limit, and which is fundamental in this work to carry out a dynamic analysis of economic aspects. Particularly, if $\{X_n : n \in \mathbb{N}\}$ is a homogeneous and finite Markov chain, with k possible states, the limiting distribution is the probability distribution given by

$$\lim_{n \rightarrow \infty} \begin{pmatrix} P_1(1) & P_2(1) & \dots & P_k(1) \end{pmatrix} \begin{pmatrix} p_{11} & p_{12} & \dots & p_{1k} \\ p_{21} & p_{22} & \dots & p_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ p_{k1} & p_{k2} & \dots & p_{kk} \end{pmatrix}^n.$$

Getting the limiting distribution in the study of this paper means that we can know what is likely to happen in the long term according to the distribution of the per capita income. Roughly speaking, the Spanish provinces tend to group into some states according to the proportions given by the limiting distribution.

However, the limiting distribution does not have to exist and, if it exists, it may not be unique, since it will depend on the initial distribution. In case the previous limit does not change for any initial distribution, we say that it is the equilibrium distribution of the chain. Furthermore, the stationary distribution is the one that does not change after each transition of the chain according to the transition probability matrix, that is, it is a row matrix p such that $pP = p$. It should be taken into account that when the equilibrium distribution exists and is unique, then it meets the stationary one.

2.1.2. States Classification

A first classification of the states of a Markov chain has to do with the access between them:

- We will say that a state x_j is accessible from another state x_i when $p_{ij} > 0$ at some instant of time. Furthermore, when it is probable to go from one state to another in both directions, we will say that both states communicate. If all the states of a Markov chain communicate, we say that the Markov chain is irreducible.
- However, if there is a state that cannot be reached from any other in the Markov chain, we will say that it is ephemeral.
- On the other hand, if there is a state from which we cannot reach any other one, we say that it is absorbing. Mathematically, the state x_i will be absorbing if $p_{ii} = 1$.

We can consider another criterion to classify the states of a Markov chain, which has to do with the probability of coming back to a certain state at some point of time. If this probability is 1, we say that the state is recurrent, while if it is less than 1, we will say that it is transitory. A well-known mathematical result establishes that two states that communicate are both recurrent or both transitory. Additionally, if we consider a recurrent state and define the random variable that describes the number of transitions needed to come back to that state, we can find its expected value, which will provide us the mean recurrence time, that is, the mean time since the state is left until the Markov chain returns to it. In particular, we talk about a positive recurrent state if its mean recurrence time is finite. On the other hand, a state is said to be periodic if, starting from it, it is only possible to return to it in a number of stages multiple of an integer greater than 1. It will be aperiodic if we can return to it after each transition. In this case, the period is 1.

Table 1 summarizes the classification of the states of a Markov chain according to both criteria.

Table 1. States classification.

Type of State	Conditions
x_j accesible from x_i	$p_{ij} > 0$
x_j ephemeral	$p_{ij} = 0$ for each i
x_i absorbing	$p_{ii} = 1$
Recurrent	Probability of coming back to it = 1
Positive recurrent	Finite mean recurrence time
Transitory	Probability of coming back to it < 1
Aperiodic	Period = 1

The next result is one of the keys of this work:

Theorem 1. *If $\{X_n : n \in \mathbb{N}\}$ is a homogeneous, finite, irreducible and aperiodic Markov chain, then the stationary distribution exists, is unique and meets the equilibrium distribution. Moreover, the mean recurrence time of each state is given by the inverse of the respective probability in the equilibrium distribution.*

This theorem gives special emphasis to homogeneous, finite and discrete-time Markov chains. It is worth noting that the software R has a package which is especially focused on the study of these stochastic processes. It is called “markovchain” and can simplify some calculations, so that we can get conclusions from a research work based on the application of dynamic probabilistic models. In order to show this methodology, we refer the reader to Appendix B.1 so that the basic codes in this software can be seen.

2.1.3. Estimating the Transition Probability Matrix

Suppose that $\{X_n : n \in \mathbb{N}\}$ is a homogeneous and finite Markov chain, with k possible states. Then the transition probability matrix is constant. If the transition probabilities are not known in advance, we can estimate them if we have data for individual transitions between two consecutive instants of time. In other words, if n_{ij} is the number of individuals that were in state x_i at time t , and reach state x_j at time $t + 1$, then the maximum likelihood estimator of the transition probability p_{ij} is given, according to [Anderson and Goodman \(1957\)](#), by

$$\hat{p}_{ij} = \frac{n_{ij}}{\sum_{j=1}^k n_{ij}}.$$

Therefore, the probability of moving from state x_i to state x_j can be calculated as the proportion of individuals who, being in the state x_i in a certain instant of time, reach the state x_j in the following period. In fact, this estimator is justified to be consistent, that is, the larger the sample size, the better the estimation made. Moreover, it is known that, although this estimator is biased, its bias decreases as the number of individuals under study increases.

2.2. Data and Treatment

Next step is choosing the data we are going to work with. We collect the GDP per capita, in euros, for the fifty Spanish provinces together with the two autonomous cities, in the time period between 2010 and 2020. We have not considered more years, since the idea is to compare the future estimation with and without COVID-19 pandemic. The source for these data is the [National Statistics Institute \(2023\)](#) (Spain). The data are collected in Table 2, in which the numbers have been rounded to three decimal places.

Table 2. GDP per capita relative to the Spanish average for each Spanish province/ autonomous city and year (2010–2020).

Province/Auton. City	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020
Almería	0.802	0.754	0.767	0.760	0.783	0.801	0.838	0.857	0.839	0.841	0.862
Cádiz	0.735	0.736	0.729	0.717	0.699	0.691	0.691	0.695	0.693	0.700	0.679
Córdoba	0.717	0.715	0.696	0.711	0.706	0.717	0.709	0.711	0.703	0.681	0.707
Granada	0.708	0.713	0.719	0.719	0.732	0.737	0.715	0.707	0.705	0.713	0.725
Huelva	0.747	0.779	0.781	0.732	0.719	0.729	0.735	0.763	0.781	0.757	0.767
Jaén	0.706	0.713	0.661	0.715	0.675	0.731	0.699	0.691	0.709	0.670	0.718

Table 2. Cont.

Province/Auton. City	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020
Málaga	0.758	0.748	0.731	0.724	0.730	0.725	0.715	0.720	0.726	0.729	0.710
Sevilla	0.814	0.815	0.818	0.799	0.803	0.790	0.780	0.782	0.783	0.785	0.795
Huesca	1.124	1.132	1.118	1.164	1.139	1.105	1.168	1.136	1.117	1.119	1.216
Teruel	1.045	1.045	1.062	1.085	1.084	1.023	0.994	0.955	0.977	0.965	0.994
Zaragoza	1.093	1.087	1.076	1.085	1.086	1.071	1.076	1.091	1.096	1.096	1.127
Asturias	0.917	0.914	0.905	0.892	0.883	0.882	0.872	0.878	0.880	0.879	0.887
Baleares	1.059	1.059	1.067	1.065	1.077	1.077	1.087	1.085	1.081	1.071	0.913
Las Palmas	0.838	0.836	0.824	0.833	0.820	0.801	0.811	0.810	0.810	0.800	0.716
Sta. Cruz de Tenerife	0.890	0.886	0.878	0.860	0.851	0.843	0.824	0.827	0.816	0.807	0.742
Cantabria	0.945	0.937	0.934	0.921	0.927	0.910	0.913	0.912	0.918	0.922	0.934
Ávila	0.788	0.801	0.818	0.808	0.802	0.788	0.775	0.776	0.784	0.791	0.828
Burgos	1.111	1.132	1.156	1.129	1.112	1.100	1.112	1.128	1.145	1.128	1.144
León	0.870	0.867	0.874	0.856	0.848	0.838	0.818	0.818	0.823	0.829	0.867
Palencia	1.022	1.043	1.024	1.028	1.010	1.027	1.062	1.004	1.054	1.041	1.069
Salamanca	0.804	0.810	0.810	0.798	0.796	0.797	0.810	0.808	0.811	0.822	0.855
Segovia	0.939	0.930	0.921	0.917	0.926	0.931	0.907	0.850	0.858	0.860	0.887
Soria	1.004	1.015	0.993	1.016	1.022	1.018	0.999	0.986	1.082	1.066	1.074
Valladolid	1.026	1.020	1.018	1.020	1.025	1.022	1.045	1.057	1.074	1.064	1.092
Zamora	0.798	0.823	0.851	0.832	0.819	0.819	0.810	0.743	0.756	0.770	0.807
Albacete	0.801	0.793	0.798	0.801	0.783	0.798	0.793	0.806	0.815	0.823	0.851
Ciudad Real	0.835	0.837	0.841	0.826	0.798	0.830	0.834	0.836	0.839	0.824	0.857
Cuenca	0.839	0.859	0.870	0.875	0.852	0.868	0.866	0.867	0.882	0.856	0.891
Guadalajara	0.832	0.836	0.830	0.814	0.770	0.742	0.757	0.776	0.789	0.794	0.816
Toledo	0.760	0.743	0.730	0.731	0.718	0.716	0.721	0.716	0.724	0.717	0.746
Barcelona	1.172	1.167	1.172	1.181	1.193	1.193	1.201	1.208	1.205	1.208	1.201
Gerona	1.156	1.143	1.151	1.144	1.149	1.143	1.154	1.093	1.080	1.084	1.083
Lérida	1.191	1.190	1.219	1.247	1.241	1.240	1.171	1.091	1.091	1.105	1.119
Tarragona	1.168	1.156	1.153	1.157	1.170	1.190	1.206	1.211	1.175	1.154	1.113
Alicante	0.769	0.749	0.737	0.738	0.749	0.747	0.758	0.762	0.756	0.754	0.762
Castellón	0.970	0.998	0.972	0.989	0.992	1.021	1.037	1.090	1.067	1.062	1.048
Valencia	0.940	0.939	0.930	0.935	0.944	0.934	0.920	0.909	0.922	0.921	0.929
Badajoz	0.716	0.709	0.694	0.702	0.690	0.702	0.704	0.715	0.711	0.704	0.738
Cáceres	0.715	0.701	0.716	0.723	0.720	0.721	0.730	0.752	0.765	0.771	0.786
La Coruña	0.949	0.938	0.931	0.940	0.924	0.935	0.941	0.930	0.940	0.935	0.952
Lugo	0.865	0.880	0.899	0.918	0.936	0.950	0.924	0.892	0.910	0.893	0.889
Orense	0.803	0.824	0.842	0.839	0.829	0.825	0.838	0.839	0.852	0.875	0.890
Pontevedra	0.856	0.842	0.840	0.852	0.856	0.854	0.850	0.871	0.859	0.869	0.904
Madrid	1.340	1.360	1.377	1.372	1.372	1.373	1.373	1.370	1.364	1.369	1.370
Murcia	0.832	0.819	0.823	0.831	0.822	0.838	0.833	0.830	0.816	0.818	0.834

Table 2. Cont.

Province/Auton. City	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020
Navarra	1.229	1.234	1.226	1.236	1.239	1.228	1.224	1.220	1.205	1.210	1.221
Álava	1.474	1.496	1.507	1.532	1.550	1.515	1.546	1.524	1.515	1.475	1.506
Vizcaya	1.235	1.227	1.239	1.231	1.245	1.246	1.237	1.212	1.217	1.221	1.220
Guipúzcoa	1.289	1.299	1.311	1.298	1.286	1.271	1.264	1.297	1.289	1.297	1.288
La Rioja	1.082	1.082	1.082	1.087	1.102	1.096	1.068	1.063	1.068	1.061	1.087
Ceuta	0.850	0.835	0.822	0.839	0.821	0.814	0.805	0.782	0.786	0.795	0.839
Melilla	0.794	0.777	0.752	0.759	0.751	0.743	0.741	0.717	0.724	0.728	0.765

The main idea is to follow the methodology proposed by Quah (1993), to analyze the evolution of GDP per capita based on data from the last decade. This analysis, with a Markovian approach, will have a double goal: evaluate the impact of COVID-19 pandemic on economic convergence by provinces in Spain; and study the speed of this convergence, both with and without COVID-19. Specifically, we start from a homogeneous and finite Markov chain, with which to study the transitions between productivity states. Productivity will be grouped into four states, “Low income”, “Medium-low income”, “Medium-high income” and “High income”, which will be given by the quartiles of the overall GDP per capita distribution of the provinces relative to the national average of each year, that is, taking as data the result of dividing the GDP per capita of each province by the Spanish GDP per capita of the corresponding year.

Next step is constructing the Markov chain transition probability matrix. With that purpose, we obtain a transition probability matrix between each pair of years of the considered period, which will give us a total of ten. To obtain each of them, the maximum likelihood estimator is used (see Section 2.1.3), according to which each transition probability can be obtained as the proportion of provinces that, being in a certain state in year t , change to a certain state in year $t + 1$. The general matrix is the result of averaging the numbers in the same position of each of the ten matrix we have constructed before. Then we check if the transition probability matrix satisfies conditions of Theorem 1, so that the stationary and equilibrium distribution can be found and meet, as well as the mean recurrence time of each state.

Finally, once we know the stationary distribution of the Markov chain, we try to get the convergence speed of the provinces towards this situation. For this purpose, Shorrocks (1978) proposes an index to analyze the mobility between states given by the transition probability matrix in which the elements of the main diagonal are each greater than or equal to the entries of the matrix that are located in the remaining positions. Specifically, this index is

$$I_1 = \frac{n - \text{tr}(\mathbf{P})}{n - 1},$$

where $\text{tr}(\mathbf{P})$ denotes the trace of the matrix $\mathbf{P}$ and n is the number of states of the Markov chain. This index gives us values between 0 and 1: mobility is null when $I_1 = 1$, because in this case $\text{tr}(\mathbf{P}) = n$ and, consequently, all states are absorbent; 0 means perfect mobility. Additionally, Sommers and Conlisk (1979) give another index with which to know the speed at which the Markov chain reaches the steady state, and it is

$$I_2 = 1 - |\lambda_2|,$$

where λ_2 is the eigenvalue of the transition probability matrix with the second highest modulus (in fact, in each Markov chain, $\lambda_1 = 1$ is always an eigenvalue, and it is the one having the greatest modulus). For further reference about measures of mobility, see, for example, [Formby et al. \(2004\)](#).

2.3. Cluster Analysis

Cluster analysis is a data analysis technique whose main goal is to group the data in a homogeneous way, which means that the elements of the same group are similar to each other in terms of the characteristic which has been analyzed, as long as the discrepancies between individuals from different groups are significant. In other words, this technique tries to minimize the intra-group variability while trying to maximize the inter-group one. In order to determine which individuals in the sample have a certain similarity, distances are generally used and, in this work, we will operate with the most classic distance: the Euclidean one. This distance, in essence, measures the longitudinal magnitude in a straight line from one point to another. There are other distances used in the construction of a cluster, such as the Manhattan distance, the Mahalanobis one or the maximum one. We will have to minimize this distance between individuals to conclude which of them are the most similar.

Individual grouping methods are classified into two groups: hierarchical and non-hierarchical clustering. In the first case, we can give a tree-based representation (called dendrogram) so that in each iteration, an order is followed and the structure to create the groups is kept. Moreover, they can be classified into two groups:

- Agglomerative: they start from simple groups which become more sophisticated as more iterations are taken. It is, therefore, an ascending approach between individuals.
- Divisive: we start from the sample as a group and, at each step, smaller groups are built until the desired number of clusters is achieved. It is, therefore, a descending approach.

In non-hierarchical clustering, the number of groups is chosen and, subsequently, individuals are included in each group, being able to move from one group to another at each step, until a certain optimality criterion is got.

In this work we will focus on the agglomerative hierarchical clustering based on Ward's method, using the Euclidean distance to find the distance between elements. The method used has to do with the way of calculating the distance between groups. It does make sense to consider the hierarchical clustering in this work, since its representation (dendrogram) is quite useful for the reader in order to have a quick idea of the relationship between provinces in terms of their per capita income. What is more, thanks to this representation, one can group provinces into a different desired number of clusters. Particularly, we have chosen the agglomerative one, which is the most common type of hierarchical clustering, indeed [Kassambara \(2017\)](#). Since we want to group provinces in terms of their per capita income, we start by treating each province as a singleton cluster and, next, pairs of clusters are successively merged until all of them belong to a single cluster, containing all provinces. Finally, it is worth noting that we have considered four clusters as recommended by the Elbow method and in order to meet the number of states in the Markov chains analysis.

For that purpose, software R is used in order to get the clustering. In order to get the final groups, we have added Appendix [B.2](#), in which the used code is explained.

3. Results and Discussion

3.1. Evolution of per Capita Income in Presence of COVID-19

Let us consider the distribution of the per capita income of the fifty provinces and two autonomous cities in Spain for all years between 2010 and 2020 relative to the annual

average. The quartiles of the overall distribution (that is, the one having $11 \times 52 = 572$ data) are the following:

- First quartile: 0.7875988. Hence, a province will be said to be in the “low income” state if its GDP per capita divided by the annual average is less than 0.7875988.
- Second quartile: 0.8672922. In case the per capita income relative to the Spanish average is greater than 0.7875988 and less than 0.8672922, the province is said to be in the “medium-low income” state.
- Third quartile: 1.0802799. In case the per capita income relative to the Spanish average is greater than 0.8672922 and less than 1.0802799, the province is said to be in the “medium-high income” state. Finally, if the per capita income relative to the Spanish average is greater than 1.0802799, the province will be said to have high income.

Hence, let us consider a discrete-time homogeneous Markov chain, $\{X_n : n \in \mathbb{N}\}$, whose possible states are given by the following four levels of income: low, medium-low, medium-high and high. The historical data from 2010 to 2020 (both years are included) let us get the following transition probability matrix, as explained in the previous section:

$$\begin{pmatrix} 0.951 & 0.049 & 0 & 0 \\ 0.072 & 0.851 & 0.077 & 0 \\ 0 & 0.061 & 0.872 & 0.067 \\ 0 & 0 & 0.06 & 0.94 \end{pmatrix}.$$

The transition diagram can be seen in Figure 1.

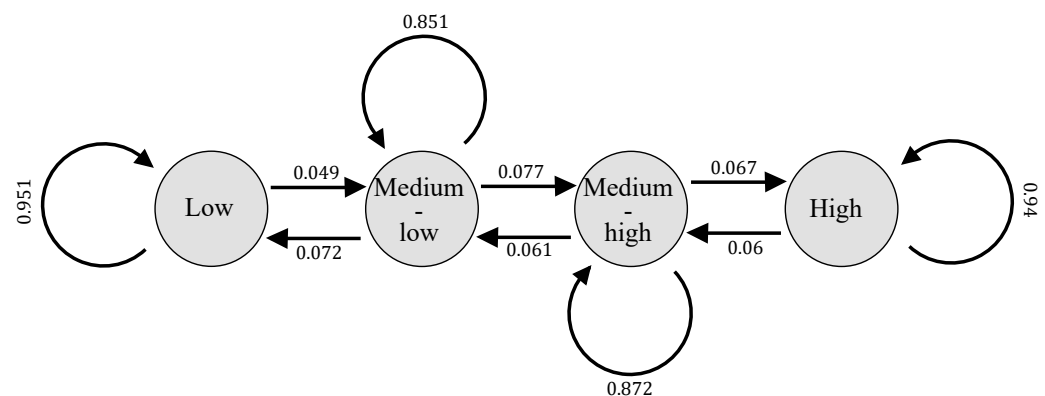


Figure 1. Transition diagram.

Each of the numbers in the previous matrix gives us the probability that provinces stay or change their income level position after a year, according to the data we have used to construct the matrix. Hence, for example, the first entry means that 95.1% of provinces that belong to the low income state, will remain in this situation next year, while the number on its right means that 4.9% of provinces which have low income in a certain year, will move on to a medium-low income level the following year.

Hence, it is clear that the Markov chain is irreducible since all the states communicate.

Once we have got the transition probability matrix, we can study some properties of the Markov chain with the help of the software R. The results are the following:

- The Markov chain is aperiodic.
- Its stationary distribution is (0.286 0.194 0.246 0.274).

Note that we have considered three decimal places in all the results, so that in the following we can conclude the percentages in the long term with just one decimal place.

Since $\{X_n : n \in \mathbb{N}\}$ is a homogeneous, finite, irreducible and aperiodic Markov chain, then the stationary distribution exists (we know it, indeed), is unique and meets

the equilibrium distribution (see Theorem 1). The stationary distribution allows us to conclude that in the long term, 28.6% of Spanish provinces will have a low income, while the percentage is 19.4% for those which will remain in a medium-low position. Additionally, 24.6% of Spanish provinces are expected to be considered as a medium-high income one in the long term, while the rest will have a high income position.

Also, this theorem lets us calculate the mean recurrence time of each state as follows:

- Low income: $1/0.286 \approx 3.5$ years.
- Medium-low income: $1/0.194 \approx 5.15$ years.
- Medium-high income: $1/0.246 \approx 4.07$ years.
- High income: $1/0.274 \approx 3.65$ years.

For example, the first mean recurrence time, 3.5 years, means that there will be, approximately, 3.5 years since provinces go out this state until they come back to that situation. Analogously, we can conclude about the other mean recurrence times.

Finally, we can calculate the speed of convergence to the stationary state and the measure of mobility between states through the two following indices:

$$I_1 = \frac{4 - 3.614}{4 - 1} = 0.129, \quad I_2 = 1 - 0.968 = 0.032.$$

The value of I_1 lets us claim that there is a low mobility between states, while the second one lets us know that the speed of convergence to the stationary distribution is quite low.

3.2. Evolution of per Capita Income Without COVID-19

In this subsection, a similar study is carried out, but considering data from 2010 to 2019, in order to avoid the effect of pandemic. The (estimated) transition probability matrix is

$$\begin{pmatrix} 0.962 & 0.038 & 0 & 0 \\ 0.063 & 0.87 & 0.067 & 0 \\ 0 & 0.05 & 0.902 & 0.048 \\ 0 & 0 & 0.065 & 0.935 \end{pmatrix}.$$

It is clear that the Markov chain is irreducible since all the states communicate. This Markov chain is also aperiodic and its stationary distribution is

$$(0.332 \ 0.201 \ 0.269 \ 0.198).$$

Since $\{X_n : n \in \mathbb{N}\}$ is a homogeneous, finite, irreducible and aperiodic Markov chain, then, by Theorem 1, the stationary distribution is unique and meets the equilibrium distribution. The stationary distribution lets us conclude that in the long term, 33.2% of Spanish provinces will have a low income, whilst the percentage is 20.1% for those which will remain in a medium-low position. Additionally, 26.9% of Spanish provinces are expected to be considered as a medium-high income one in the long term, while the rest will have a high income position.

Also, this theorem lets us calculate the mean recurrence time of each state as the inverse of each probability in the stationary distribution:

- Low income: $1/0.332 \approx 3.01$ years.
- Medium-low income: $1/0.201 \approx 4.98$ years.
- Medium-high income: $1/0.269 \approx 3.72$ years.
- High income: $1/0.198 \approx 5.05$ years.

Hence, for example, provinces which leave the low income state, will spend, on average, 3.01 years until they come back to that situation.

Moreover, we can claim that the mobility between states for the Spanish provinces in terms of per capita income and the speed of convergence towards the stationary state are very reduced, since

$$I_1 = \frac{4 - 3.669}{4 - 1} = 0.11 \quad \text{and} \quad I_2 = 1 - 0.971 = 0.029.$$

If we compare these two values with those obtained in the previous subsection, we can note that the speed of convergence to the steady state is slightly higher in the current scenario (with COVID-19). What is more, the mobility between different states is also higher in this situation.

On the other hand, it makes sense to estimate the distribution of provinces by states in 2020 by using the Markov chain in this subsection. Hence, we will be able to compare the results with those given by the [National Statistics Institute \(2023\)](#) database. According to Table 2, and taking into account the 2010–2020 time period, in 2020 there were 13 provinces with low income, 11 with medium-low income, 14 with medium-high income and 14 with high income. Thanks to the study we have carried out in this subsection, we can find the prediction for 2020 based on the transition probability matrix obtained. In 2019, 13 provinces had low income, 13 had medium-low income, 14 had medium-high income and 12 had high income. Thus, according to the Markov model, if the pandemic had not have existed, the distribution of provinces in each income state in 2020 would have been given by

$$(0.25 \ 0.25 \ 0.269 \ 0.231) \begin{pmatrix} 0.962 & 0.038 & 0 & 0 \\ 0.063 & 0.87 & 0.067 & 0 \\ 0 & 0.05 & 0.902 & 0.048 \\ 0 & 0 & 0.065 & 0.935 \end{pmatrix} = (0.256 \ 0.241 \ 0.274 \ 0.229).$$

This means that the number of provinces in each state is, respectively,

- $0.256 \times 52 \approx 13.31$.
- $0.241 \times 52 \approx 12.53$.
- $0.274 \times 52 \approx 14.25$.
- $0.229 \times 52 \approx 11.91$.

We conclude that, without COVID-19, the number of provinces in the medium-low income state would have been greater, while the number of provinces in the high income state would have been smaller. The other states would have been stable. With this prediction, it is worth noting the role of the pandemic in mitigating the divergence between provinces.

3.3. Convergence Clubs

One of the main advantages of the Markov chain analysis is that it lets us conclude about the future behavior of the economic variable we have considered. However, while it is true that we have analyzed the evolution of per capita income in Spain by provinces, we still do not have an idea of which provinces belong to each state in the present and will do in the future. One way to make ourselves an idea about the groups of income they form is through a cluster analysis, which is carried out in this part of this work, as detailed in Section 2.3.

In Figure 2, each number in the x-axis refers to the row where each province is located in the Table 2, so that we can identify them. Moreover, the numbers in the y-axis have to do with the (Euclidean) distance between the observations (provinces) according to their

income level, so that the lower the distance level, the closer the observations are in each cluster, which also means that the provinces in it have a similar level of income.

If we are interested in obtaining four groups, the result, by relating each number located below in Figure 2 with the position occupied by the province in the database used (see Table 2), is as follows:

- Group 1 (high income): Navarra, Vizcaya, Lérida, Barcelona, Tarragona, Álava, Madrid and Guipúzcoa.
- Group 2 (medium-high income): Zaragoza, La Rioja, Gerona, Huesca, Burgos, Castellón, Soria, Palencia, Valladolid, Teruel and Baleares Islands.
- Group 3 (medium-low income): La Coruña, Cantabria, Valencia, Lugo, Asturias, Segovia, Orense, Cuenca, Pontevedra, León, Ciudad Real, Murcia, Las Palmas, Santa Cruz de Tenerife, Guadalajara, Sevilla, Ávila, Zamora, Ceuta, Almería, Salamanca and Albacete.
- Group 4 (low income): Cádiz, Jaén, Granada, Córdoba, Badajoz, Málaga, Toledo, Cáceres, Melilla, Huelva and Alicante.

Thus, we might talk about four groups of provinces that, after the last decade, have a similar income level and that, possibly, continue as a convergence club in terms of productivity.

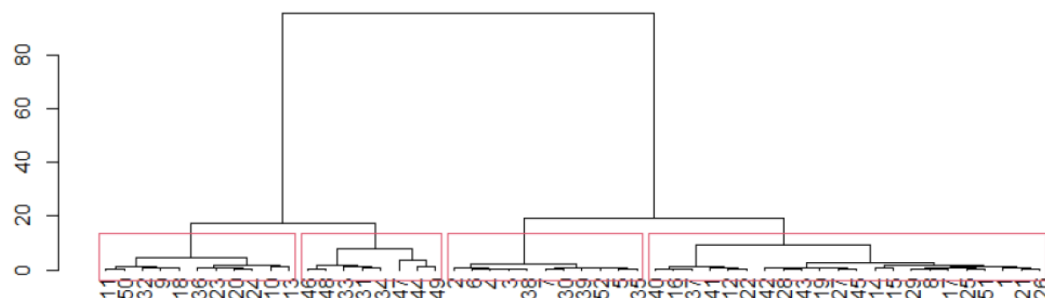


Figure 2. Dendrogram for the cluster analysis (agglomerative hierarchical based on Ward's method using the Euclidean distance).

4. Conclusions

In what follows, we highlight the main findings of this work. It can be seen that, in the long term, the Spanish provinces are expected to be grouped in different productivity states, depending on the proportions obtained. Specifically, it is estimated that approximately 28.6% of the provinces will have low income, while around 19.4% will be characterized by medium-low income. Furthermore, 24.6% will exhibit medium-high income and the remaining 27.4% are expected to achieve high income. On the other hand, it has been predicted that the provinces that leave the state of low income will require an average of three and a half years to return to that state. The mean recurrence time is, approximately, five years for provinces with medium-low income, four years for those with medium-high income and about three and a half years for those with high income. Therefore, the tendency is different for the one concluded by Gardeazábal (1996) for the period between 1967 and 1991. Furthermore, the convergence indices allow us to conclude that there is a low mobility between states of the provinces, at the same time that the convergence process is slow towards the stationary state. We also obtain that, in a hypothetical situation of absence of a pandemic, in the long term, it is estimated that 33.2% of the provinces will have low income. Moreover, in this situation, 20.1% will be characterized by having medium-low income, while 26.9% will show medium-high income. Finally, it is expected that the remaining 19.8% achieve high income. In other words, a tendency is observed for the Spanish provinces to group together in the productivity states studied, which has been influenced by the pandemic, with a notable change in extreme income levels. Although in

the current situation, 28.6% of provinces are destined to have a low level of income, in a situation without COVID-19 this percentage would be 33.2%. Hence, the pandemic has reduced the expected percentage of provinces in low-income status in the long term. What is more, if 27.4% are expected to have a high long-term income, this percentage would have been 19.8% if the pandemic had not existed. Thus, the health emergency has influenced the long-term behavior of the Spanish economy, increasing the percentage of provinces that, in the long term, will be at a high income level. Hence, it is worth noting that the pandemic has implied that the differences between provinces (in terms of income) have been mitigated. Moreover, in a situation without a pandemic, the speed of convergence is slower than in the presence of COVID-19, which confirms the role of the pandemic as a shock absorber of income differences. On the other hand, although Markov analysis does not allow us to know which provinces will be in each state in the long term, we can generate groups with provinces that have similar income levels based on data from the last decade through a cluster analysis. This allows us to get an idea about possible convergence clubs, that is, we can distinguish four groups of provinces that, at the end of the last decade, present similar levels of income and are likely to remain as convergence clubs in terms of income in the near future. Hence, the main limitation of this study is the fact that it is not possible to determine which provinces will be in each group of level of income in the long term. However, we can have an idea of their distribution, by levels of income, and compare it to the case in which the pandemic had not existed. Also, as stated before, this lack of information is mitigated by considering a cluster analysis. Finally, we want the reader to consider this work in order to know the basic commands in R to study the evolution of a variable through a Markov chain, or to group the individuals according to that variable through a cluster analysis.

Author Contributions: Conceptualization, J.F.G.-R. and A.V.G.-L.; Investigation, M.M.-H.; Methodology, J.F.G.-R., M.M.-H. and A.V.G.-L.; software, J.F.G.-R., M.M.-H. and A.V.G.-L.; Supervision, J.F.G.-R. and A.V.G.-L.; data curation, J.F.G.-R., M.M.-H. and A.V.G.-L.; Writing—original draft, M.M.-H.; Writing—review & editing, J.F.G.-R. and A.V.G.-L. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Informed Consent Statement: Not applicable.

Data Availability Statement: In order to know the GDP per capita data of all Spanish provinces see the link in the reference [National Statistics Institute \(2023\)](#).

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

Abbreviations

The following abbreviations are used in this manuscript:

GDP Gross domestic product

Appendix A. Preliminaries on Markov Chains

A stochastic process is an arbitrary collection of random variables $\{X_t : t \in T\}$, defined on the same probability space, where t is an index which usually refers to time and T is the set of indices, known as the parametric space. Depending on whether the parametric space is countable or not, it is said to be discrete or continuous, respectively. In the first case, the stochastic process is also known as a time series. Moreover, the possible values of the random variables are called states, and they are collected in the state space, which we will denote by E . If E is countable, it is said to be discrete, and the stochastic process

is called a chain; while if it is not countable, the state space is continuous. A chain whose state space is finite is said to be finite.

A discrete-time Markov chain is a stochastic process on the form $\{X_n : n \in \mathbb{N}\}$ (where $\mathbb{N} = \{1, 2, 3, \dots\}$ denotes, as usual, the set of natural numbers), whose state space is countable, and in which the value of the next variable depends only on the value of the current variable, and not on any variables in the past, known as the Markov property. That is, for each natural $n > 1$ and each $x_1, x_2, \dots, x_n$ in the state space, it holds that

$$P[X_n = x_n | X_{n-1} = x_{n-1}, \dots, X_2 = x_2, X_1 = x_1] = P[X_n = x_n | X_{n-1} = x_{n-1}].$$

Note that this property lets us claim that the future only depends on the past by taking into account the present. The fact that we deal with discrete-time has to do with the data we use in the study; since several years give us a countable set of index, the Markov is required to be discrete-time. For the pair of states x_i and x_j , we denote by p_{ij} the probability of moving from state x_i in time n to state x_j in time $n + 1$, that is,

$$p_{ij} = P[X_{n+1} = x_j | X_n = x_i].$$

It is known as transition probability and, in case this value is the same for each n , the Markov chain is said to be homogeneous.

Suppose that $\{X_n : n \in \mathbb{N}\}$ is a finite and homogeneous discrete-time Markov chain, such that the range of each random variable is enumerated according to the indices of the set $\{1, \dots, k\}$. Then we can collect the transition probabilities in a matrix,

$$P = \begin{pmatrix} p_{11} & p_{12} & \dots & p_{1k} \\ p_{21} & p_{22} & \dots & p_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ p_{k1} & p_{k2} & \dots & p_{kk} \end{pmatrix},$$

called the transition probability matrix and its terms satisfy the following conditions:

- $p_{ij} \geq 0$ for each $i, j \in \{1, \dots, k\}$.
- $\sum_{j=1}^k p_{ij} = 1$ for each $i \in \{1, \dots, k\}$.

Note that P is a square matrix and each of its rows is a discrete probability distribution according to the previous properties. In fact, the transition probability matrix is said to be a stochastic matrix, precisely because it satisfies these properties. Also, note that P^n is a stochastic matrix for each natural number n .

On the other hand, the probability distribution from which the chain starts, and which is given by $P[X_1 = x_j] = P_j(1)$ for each $k = 1, \dots, n$, which is called the initial distribution, together with the transition probability matrix, determine the probability distribution of the Markov chain. Indeed, we can obtain the probability distribution at any moment of time by multiplying the initial distribution by the transition probability matrix as many times as required. In general, we can write

$$\begin{pmatrix} P_1(n+1) & P_2(n+1) & \dots & P_k(n+1) \end{pmatrix} = \begin{pmatrix} P_1(1) & P_2(1) & \dots & P_k(1) \end{pmatrix} P^n$$

for each $n \in \mathbb{N}$.

Appendix B. Procedures in R

Appendix B.1. Markov Chains

Once we have got the transition probability matrix, if we denote it by `mc`, and then load the package (`library(markovchain)`), we can get the stationary distribution through

```
steadyStates(mc)
```

Moreover, we can check if it is aperiodic by using the command

```
period(mc)
```

If the result is 1, the Markov chain is aperiodic. We can also get the mean recurrence time of each state:

```
meanRecurrenceTime(mc)
```

Finally, finding out if it is irreducible is possible by writing

```
is.irreducible(mc)
```

and checking if the result is “TRUE” or “FALSE”. As a complement, the command

```
plot(mc)
```

gives us the transition diagram, which is a plot where we can see the connection between states through the transition probabilities.

Appendix B.2. Cluster Analysis

Next, we expose the steps that have been followed in order to get the final groups when clustering (agglomerative hierarchical clustering) provinces according to their income in the period 2010–2020:

1. Standardize the variables, that is, subtract the mean from the value of each one and divide the result by the standard deviation of the values of the variable. If the data contain the information to be processed, we must implement

```
df=as.data.frame(scale(data))
```

2. Calculate the proximity matrix by using the Euclidean distance:

```
d_eu <-dist(df, method = 'euclidean' )
```

3. Find the agglomerative hierarchical cluster with Ward's method:

```
cluster <- hclust(d_eu, method = 'ward.D')
```

4. Draw the dendrogram:

```
plot(as.dendrogram(cluster))
```

5. Draw rectangles that group a certain number, k, of the individuals in the sample:

```
rect.hclust(cluster, k = 4)
```

References

- Anderson, T. W., & Goodman, L. A. (1957). Statistical inference about Markov chains. *The Annals of Mathematical Statistics*, 28(1), 89–110. [\[CrossRef\]](#)
- Arreola, D., & Montiel, L. V. (2024). Approximating income inequality dynamics given incomplete information: An upturned Markov chain model. *Computational Statistics*, 39(2), 629–651. [\[CrossRef\]](#)
- Ayuda, M. I., Collantes, F., & Pinilla, V. (2010). Long-run regional population disparities in Europe during modern economic growth: A case study of Spain. *The Annals of Regional Science*, 44, 273–295. [\[CrossRef\]](#)
- Bode, E., & Nunnenkamp, P. (2011). Does foreign direct investment promote regional development in developed countries? A Markov chain approach for US states. *Review of World Economics*, 147, 351–383. [\[CrossRef\]](#)
- Chen, Y., Mamon, R., Spagnolo, F., & Spagnolo, N. (2022). Renewable energy and economic growth: A Markov-switching approach. *Energy*, 244, 123089. [\[CrossRef\]](#)

- Deaton, A. (2021). COVID-19 and global income inequality. *LSE Public Policy Review*, 1(4), 1. [CrossRef]
- Everitt, B., Landau, S., Leese, M., & Stahl, D. (2011). *Cluster Analysis* (5th ed.). John Wiley & Sons.
- Fingleton, B. (1997). Specification and testing of Markov chain models: An application to convergence in the European Union. *Oxford Bulletin of Economics and Statistics*, 59(3), 385–403. [CrossRef]
- Formby, J. P., Smith, W. J., & Zheng, B. (2004). Mobility measurement, transition matrices and statistical inference. *Journal of Econometrics*, 120(1), 181–205. [CrossRef]
- Gardeazábal, J. (1996). Provincial income distribution dynamics: Spain 1967–1991. *Investigaciones Económicas*, 20(2), 263–269.
- Gostkowski, M., Rokicki, T., Ochnio, L., Koszela, G., Wojtczuk, K., Ratajczak, M., Szczepaniuk, H., Bórawski, P., & Beldycka-Bórawska, A. (2021). Clustering analysis of energy consumption in the countries of the visegrad group. *Energies*, 14(18), 5612. [CrossRef]
- Haller, A., Gherasim, O., & Bălan, M. (2020). Medium-term forecast of European economic sustainable growth using Markov chains. *Zb. rad. Ekon. fak. Rij.*, 38(2), 585–618.
- He, L., Tao, J. G., Meng, P., Chen, D., Yan, M., & Vasa, L. (2021). Analysis of socio-economic spatial structure of urban agglomeration in China based on spatial gradient and clustering. *Oeconomia Copernicana*, 12(3), 789–819. [CrossRef]
- Karahasan, B. C. (2020). Can neighbor regions shape club convergence? Spatial Markov chain analysis for Turkey. *Letters in Spatial and Resource Sciences*, 13(2), 117–131. [CrossRef]
- Kassambara, A. (2017). *Practical guide to cluster analysis in R: Unsupervised machine learning*. Sthda.
- Kerkouch, A., Bensbahou, A., Seyagh, I., & Agouram, J. (2024). Dynamic analysis of income disparities in Africa: Spatial markov chains approach. *Scientific African*, 24, e02236. [CrossRef]
- Le Gallo, J., & Chasco, C. (2008). Spatial analysis of urban growth in Spain, 1900–2001. *Empirical Economics*, 34, 59–80. [CrossRef]
- Lipták, K. (2011). The application of Markov chain model to the description of hungarian market processes. *Zarządzanie Publiczne*, 16(4), 133–149.
- National Statistics Institute. (2023). *Regional Accounting of Spain. Results. GDP and GDP per Capita. 2000–2022 Series*. Available online: https://www.ine.es/dyngs/INEbase/es/operacion.htm?c=Estadistica_C&cid=1254736167628&menu=resultados&idp=1254735576581 (accessed on 2 October 2023).
- Papanikolaou, N. (2020). Markov-switching model of family income quintile shares. *Atlantic Economic Journal*, 48(2), 207–222. [CrossRef]
- Quah, D. (1993). Empirical cross-section dynamics in economic growth. *European Economic Review*, 37, 426–434. [CrossRef]
- Rey, S. (2023). Intersectional urban dynamics: A joint Markov chains approach. *Letters in Spatial and Resource Sciences*, 16(1), 36. [CrossRef]
- Shorrocks, A. F. (1978). The measurement of mobility. *Econometrica: Journal of the Econometric Society*, 46(5), 1013–1024. [CrossRef]
- Sommers, P. S., & Conlisk, J. (1979). Eigenvalue immobility measures for Markov chains. *Journal of Mathematical Sociology*, 6, 253–276. [CrossRef]
- Takács, L. (1960). *Stochastic processes problems and solutions*. Chapman and Hall.
- Tirado, D. A., Díez-Minguela, A., & Martínez-Galarraga, J. (2016). Regional inequality and economic development in Spain, 1860–2010. *Journal of Historical Geography*, 54, 87–98. [CrossRef]
- Wenxuan, Y. (2023). Human capital dynamics across provinces in china: A spatial markov chain approach. *Forum of International Development Studies*, 53(8), 1–17.
- Yang, Y., & Hu, A. (2008). Investigating regional disparities of China's human development with cluster analysis: A historical perspective. *Social Indicators Research*, 86, 417–432. [CrossRef]
- Zarikas, V., Pouloupoulos, S. G., Gareiou, Z., & Zervas, E. (2020). Clustering analysis of countries using the COVID-19 cases dataset. *Data in Brief*, 31, 105787. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.