


Article

Research on the Self-Repairing Model of Outliers in Energy Data Based on Regional Convergence

Nan Li ^{1,*}, Xunwen Zhao ¹ , Hailin Mu ¹, Yimeng Li ¹, Jingru Pang ¹, Yuqing Jiang ², Xin Jin ¹ and Zhenwei Pei ¹

¹ Key Laboratory of Ocean Energy Utilization and Energy Conservation of Ministry of Education, Dalian University of Technology, Dalian 116024, China; zhaoxw666@mail.dlut.edu.cn (X.Z.); hailinmu@dlut.edu.cn (H.M.); li_yimeng@mail.dlut.edu.cn (Y.L.); pjr@mail.dlut.edu.cn (J.P.); jinxin1123@mail.dlut.edu.cn (X.J.); peizhenwei@mail.dlut.edu.cn (Z.P.)

² School of Economics and Management, China University of Petroleum, Beijing 102249, China; 2019310906@student.cup.edu.cn

* Correspondence: nanli.energy@dlut.edu.cn; Tel.: +0411-84707057

Received: 3 August 2020; Accepted: 15 September 2020; Published: 18 September 2020



Abstract: The need for the statistical stability of data is increasing nowadays as the data resource has become a more and more important production factor. In this study, a set of general identification and correction models are established for data outlier modification. The research object we chose is the data of per capita energy consumption. Based on the joint diagnosis method of outliers and the regional convergence theory, the abrupt outliers are identified and corrected. The study finds that there is an outlier in the data of the Ningxia Hui Autonomous Region. According to the club grouping method, 30 provinces in China are divided into two clubs and the Ningxia Hui Autonomous Region is determined to be in the first club. We calculate the convergence rate and obtain the correction results combining the half-life cycle model.

Keywords: club convergence; outliers; energy consumption; half-life cycle; time series

1. Introduction

With the rapid development of information technology, the data resource has become an important factor of production, with more attention being paid to its stability. The energy industry is the foundation of national economic and social development, which is also deeply influenced by big data. Energy statistics are the cornerstone of energy research and policymaking. People have been trying to make the energy data as complete as possible, and few attempts have been made to verify and correct abnormal data. In this paper, we utilize regional convergence theory and the half-life cycle model to correct outlier in energy statistic data. For the identification process of outliers, the joint diagnosis method based on the autoregressive moving average (ARMA) model of a time series is used.

The study of regional convergence theory began in the mid-20th century, developed from Solow's neoclassical growth model [1]. The main idea is that economies tend to develop into the same steady-state from time to time due to the nature of diminishing marginal returns. In the 1980s, the study on convergence began to develop and the types of convergence were divided into four main types: σ convergence, β convergence, γ convergence and club convergence. Baumol (1986) [2] conducted an empirical study on convergence first and showed that the growth rate of productivity in industrialized countries is negatively corrected with its productivity level. Barro and Sala-i-Martin (1991) [3] proposed the econometric definitions of sigma convergence and beta convergence. Sigma convergence refers to the narrowing of the average income gap over time. Beta convergence means the growth rate of per capita income in underdeveloped regions is higher than that in developed regions, which shows a

“catch-up effect”. Barro (2003) [4] analyzed panel data of per capita GDP of 100 countries from 1960 to 1990 and verified the existence of conditional convergence. Chambers (2016) [5] constructed the income distribution convergence model of 81 countries from 1990 to 2010 using the Gini coefficient and income mean logarithmic deviation cross-section data and panel data. Leonardo (2018) [6] studied whether the labor productivity of different regions in Peru convergent from 2002 to 2012, the results show that the labor productivity of Peru’s secondary industry (especially manufacturing industry) is convergent, while the labor productivity of agriculture and service industry is not convergent. With regard to the development of regional convergence theory, the research field gradually extended from economics to others, and an increasing amount of papers begin to study the convergence research of energy issues. List (1999) [7] used the time-series analysis method for the first time to study whether the income convergence in the United States is accompanied by the convergence of air pollutant emissions from 1929 to 1994. Mishra (2014) [8] uses panel Kwiatkowski–Phillips–Schmidt–Shin (KPSS) stationarity test and Lagrange multiplier unit root test to verify the convergence of energy consumption per capita in Association of Southeast Asian Nations (ASEAN) countries from 1971 to 2011. Sheng (2014) [9] analyzed the panel data of 27 provinces in China from 1978 to 2008, and used regression to study the relationship between China’s economic growth, energy demand and relevant policies. Apergis (2016) [10] tested the convergence of wholesale electricity prices in big Australian states. Solarin (2018) [11] used a more flexible fractional integrator-based method to test the hypothesis of convergence of renewable energy consumption in 27 Organization for Economic Co-operation and Development (OECD) countries.

In the mid-1990s, with the development of the study on time and space characteristics of the regional economy, club convergence theory began to get more attention. Club convergence refers to the convergence of economic growth of a group of regions in the same initial conditions and structural characteristics of economic growth to the same steady state. Barro and Sala-i-Martin (1991) propose the concept of club convergence, and they regard club convergence as the β convergence of a regional group that has similar initial economic conditions and structural characteristics of economic growth. Galor (1996) [12] argued that the traditional neoclassical growth model gives rise to the club convergence hypothesis and conditional convergence hypothesis, and also proposes the concept of club convergence. Phillips and Sul (2007) [13,14] propose the nonlinear time-varying factor model(log- t), which is an important method of convergence test based on regression. Kim (2012) [15] studied the dynamic behavior of electricity consumption and used the log- t test to conduct club grouping and convergence tests for 109 countries from 1971 to 2009. Paker (2016) [16] analyzed the convergence of global economic and manufacturing energy productivity clubs by taking 33 countries including OECD and non-OECD nations as samples. Based on the log- t method, there are energy productivity clubs and six manufacturing clubs in the global economy. Adrian (2016) [17] tested the homogeneity of 12 tourist source markets in Spain. Cuñado [18–20] examined the real convergence hypothesis in some countries by means of using time-series techniques.

The half-life model is established by Art Schneiderman in the United States, and its main content is that the failure rate level presents a negative correlation over a certain period. When giving the defect rate at the initial level and at the lowest level, the time required to reduce the defect rate by half is the half-life cycle. Romer (2001) [21] pointed out that the half-life cycle can be calculated when the convergence rate is known, which is the number of years needed to reduce the national income gap by half. Moon (2017) [22] investigated the half-life cycle of the relative consumer price index in 15 regions of Korea from 1990 to 2016. Bergman (2017) [23] analyzed the half-life circle of 124 homogeneous commodities in Denmark from 1997 to 2010, which was 9.4 months when taking deviation into account, and 28.7 months when adopting the standard method.

The diagnosis of time-series outliers began in the 1970s. Fox (1972) [24] was the first one to define and classify outliers, who took two kinds of outliers into account in the time-series model: additive outliers (AO) that affect only one observation, and innovative outliers (IO) that continuously affect several observations. Tsay (1986) [25] further divided the outliers into three types: mean drift (LS), temporary change (TC) and variance change (VC). Bruce and Martin (1989) [26] studied the

identification method of structural changes in time series, namely taking the ARMA model as the basic framework. Chen and Liu (1993) [27] proposed a joint estimation method for the identification of outliers. Based on the Bayesian framework, Chen (2015) [28] proposed a method to comprehensively detect the location of outliers in the time-series model, and they demonstrate the effectiveness of the method through simulation research. Yan (2018) [29] proposed a method based on the Gaussian process prediction model to detect outliers in time-series data with time-varying disturbances. Sathe (2018) [30] proposed a simple subspace outlier detection algorithm; the results show that the method is much faster and the accuracy of detection is higher than traditional algorithms.

This paper introduces the theory of regional convergence and the concept of big data into the study of the energy model, combining the time-series analysis method and the half-life cycle method to establish a methodology system of self-correction of energy mutation data. The paper is mainly composed of the following aspects: the second chapter introduces the main research methods of this paper, the third chapter shows the data we use in the study, the fourth chapter analyzes the research results, and the fifth chapter is the conclusion of this paper.

2. Methodology

This paper uses a joint estimation method to identify outliers and regional convergence combining the half-life cycle method to correct outliers.

2.1. Identification of Outliers

2.1.1. Definition and Classification of Outliers

Time-series analysis refers to analyzing time series systematically and establishing reasonable models. Outliers often appear in time series, which have a serious impact on the application of data analysis. Outliers are usually the points that deviate from other observation data and therefore are obviously different from others. The first definition and classification of outliers are pointed out by Fox (1972), who divided data outliers into AO and IO. AO affects only one observation, which means that the time series will return to the normal path after passing through it, while IO affects several data points continuously.

2.1.2. ARMA Model

In order to avoid the biases in energy statistics, it is necessary to identify the outliers. The identification process is based on the ARMA model of time series. The ARMA model, known as the autoregressive moving average model, can be subdivided into three categories: autoregressive (AR) model, moving average (MA) model and ARMA model.

The structure of the ARMA (p, q) model is as follows:

$$\begin{aligned} x_t &= \varphi_0 + \varphi_1 x_{t-1} + \varphi_2 x_{t-2} + \dots + \varphi_p x_{t-p} + \varepsilon_t - \theta_1 \varepsilon_{t-1} - \theta_2 \varepsilon_{t-2} - \dots - \theta_q \varepsilon_{t-q} \\ \varphi_p &\neq 0, \theta_q \neq 0 \\ E(\varepsilon_t) &= 0, \text{Var}(\varepsilon_t) = \sigma_\varepsilon^2, E(\varepsilon_t \varepsilon_s) = 0, s \neq t \\ E(\varepsilon_s \varepsilon_t) &= 0, \forall s < t \end{aligned} \quad (1)$$

when $\varphi_0 = 0$, delay operator B is introduced:

$$\phi(B)X_t = \Theta(B)\varepsilon_t \quad (2)$$

where $\phi(B) = 1 - \varphi_1 B - \varphi_2 B^2 - \dots - \varphi_p B^p$ is the polynomial of p -order autoregression coefficient $\Theta(B) = 1 - \theta_1 B - \theta_2 B^2 - \dots - \theta_q B^q$ is the polynomial of q -order moving average coefficient.

2.1.3. Outliers Joint Estimation Method

Assuming $\{Y_t\}$ is a stationary time series and follows the ARMA process, its definition is:

$$Y_t = \frac{\theta(B)}{\alpha(B)\phi(B)}a_t \quad t = 1, \dots, n, \quad (3)$$

where n is the number of observed values of the sequence.

Considering that time series are affected by non-repeatable events, the model becomes:

$$Y_t^* = Y_t + \omega \frac{A(B)}{G(B)H(B)}I_t(t_1) \quad (4)$$

where Y_t is the same as the above equation, when $t = t_1$, $I_t(t_1) = 1$; in other cases, $I_t(t_1) = 0$. $I_t(t_1)$ is the indicator showing whether the outliers appear or not. It can locate the location where outliers appear. $A(B)/\{G(B)H(B)\}$ reveals the type of outliers.

$$\text{IO} : \frac{A(B)}{G(B)H(B)} = \frac{\theta(B)}{\alpha(B)\phi(B)} \quad (5)$$

$$\text{AO} : \frac{A(B)}{G(B)H(B)} = 1 \quad (6)$$

Y_t is supposed that there are m different types of outliers in time series, then the model containing all outliers is expressed as:

$$Y_t = \sum_{j=1}^m \omega_j L_j(B) I_t(t_j) + \frac{\theta(B)}{\alpha(B)\phi(B)}a_t \quad (7)$$

which is called joint estimation diagnostic model.

The residual sequence is:

$$\hat{e}_t = \sum_{j=1}^m \omega_j \pi(B) L_j(B) I_t(t_j) + a_t \quad (8)$$

where

$$\begin{aligned} \hat{\omega}_{IO}(t_1) &= \hat{e}_{t_1} & \hat{\tau}_{IO}(t_1) &= \hat{\omega}_{IO}(t_1) / \hat{\sigma}_a \\ \hat{\omega}_{AO}(t_1) &= \frac{\sum_{t=t_1}^n \hat{e}_t x_{2t}}{\sum_{t=t_1}^n x_{2t}^2} & \hat{\tau}_{AO}(t_1) &= \{\hat{\omega}_{AO}(t_1) / \hat{\sigma}_a\} \left(\sum_{t=t_1}^n x_{2t}^2 \right)^{1/2} \end{aligned} \quad (9)$$

2.1.4. Outliers Identification Process

After the establishment of the time-series ARMA model, the identification process of outliers can be further carried out.

The first stage: preliminary parameter estimation and outlier test.

Obtain residuals based on original or adjusted sequences and calculate model parameters of time series through maximum likelihood estimation.

For $t = 1, \dots, n$, according to the residuals obtained by 1.1, calculate $\hat{\tau}_{IO}(t)$, $\hat{\tau}_{AO}(t)$, $\eta_t = \max\{|\hat{\tau}_{IO}(t)|, |\hat{\tau}_{AO}(t)|\}$ respectively. C is the critical value, if $\max_t \eta_t = |\hat{\tau}_{tp}(t_1)| > C$, then there may be an outlier at time t_1 , whose type may be IO or AO.

If no outlier is found, go to step 1.4. Otherwise, remove the identified outliers from the sequence and return to 1.2 to check additional outliers.

If there is no outlier in the first iteration process, it shows the original sequence is not affected by outliers. The set of identified outliers go to the second stage.

The second stage: joint estimation of outlier effect and model parameters.

Suppose there are m outliers of various types, which are located at t_1, t_2, \dots, t_m . ω_j is estimated by multiple regression using Equation (8).

Calculate the $\hat{\tau}_j$ statistic by $\hat{\tau}_j = \hat{\omega}_j / \text{std}(\hat{\omega}_j)$, $j = 1, \dots, m$. If $\min_j |\hat{\tau}_j| = \hat{\tau}_v \leq c$ (c is the critical value in 1.2), then delete the outlier at t_v in m outliers. The remaining $m - 1$ outliers are returned to 2.1 to continue the iteration. Otherwise, enter 2.3.

The set of outliers obtained through calculation is considered to be real outliers of time series. Remove them and return to 1.1 to estimate model parameters and calculate the residual sequence.

2.2. Outlier Correction

2.2.1. Regional Convergence Theory

Unconditional β Convergence

Unconditional β convergence refers to the fact that the growth rate of per capita income of the underdeveloped region is higher than that of developed regions. Therefore, the lower the economic development level, the higher the growth rate of per capita income; and developing regions show a “catch-up effect” compared with developed regions. The unconditional β convergence model is expressed as follows:

$$\frac{1}{t-t_0} \ln\left(\frac{Y_{i,t}}{Y_{i,t_0}}\right) = \alpha - \left[\frac{1-e^{-\beta t}}{t-t_0}\right] \cdot \ln(Y_{i,t_0}) + \varepsilon_{i,t} \quad (10)$$

where i represents the region, t is the year, t_0 is the starting year. In this paper $\frac{1}{t-t_0} \ln\left(\frac{Y_{i,t}}{Y_{i,t_0}}\right)$ means the average growth rate of energy consumption per capita; β is the rate of convergence. The sign of β is used to measure the existence of unconditional β convergence. When $\beta > 0$, it means the energy consumption per capita in China is converged, conversely when $\beta < 0$ means there is no convergence.

Conditional β Convergence

Conditional β convergence theory adds conditional quantity as an influencing factor under the model of unconditional convergence, the model is:

$$\frac{1}{t-t_0} \ln\left(\frac{IE_{t,i}}{IE_{0,i}}\right) = \alpha - \left[\frac{1-e^{-\beta t}}{t-t_0}\right] \cdot \ln(IE_{0,i}) + c_i IF_i + u_{i,t} \quad (11)$$

where each indicator has the same meaning as the above equation. IF_i is the influencing factor of convergence, c_i is the coefficient of influencing factor.

2.2.2. Club Convergence

Club convergence means that with similar initial conditions and structural characteristics, the economic growth of a group of regions will converge to the same steady state. The regions in the same club have homogeneity, so it is of practical significance to examine the club's convergence. Grouping by club convergence theory can make a more reasonable analysis of the similarities among different provinces within the club. This paper utilizes a log- t method to study club convergence.

Nonlinear Time-Varying Factor Model

Phillips and Sul (2007) proposed a nonlinear single-factor model:

$$\begin{aligned} X_{it} &= \delta_{it} \mu_t, \\ \delta_{it} &= \delta_t + \sigma_i \xi_{it} L(t)^{-1} t^{-\alpha} \end{aligned} \quad (12)$$

where δ_i is fixed, δ_{it} is independent identically distributed in cross-section on time series. $L(t)$ is a slowly varying function, as $t \rightarrow \infty$, $L(t) \rightarrow \infty$.

Log-T Regression

Phillips and Sul proposed a simple method for detecting convergence. The steps are as follows:

- (1) calculate the cross-sectional variance ratio H_1/H_t
- (2) regression:

$$\log\left(\frac{H_1}{H_t}\right) - 2\log L(t) = \hat{a} + b \log t + \hat{u}_t \quad (13)$$

In the Equation, $t = [rT], [rT] + 1, \dots, T$, T is the time span, $r = 0.3$ according to recommendation. Slowly changing function $L(t) = \log(t+1)$. HACt statistic $t_{\hat{b}}$ of \hat{b} is obtained by OLS regression.

Club Grouping

Based on the log- t method, Phillips and Sul (2007) proposed an endogenous identification algorithm for convergence club. The specific steps are as follows:

Sort: the panel data are sorted from large to small according to the cross-section data of last year.

- (1) Form core group: calculate from the section element with the first order, add one element at a time in log- t regression. The calculated value $t_{\hat{b}}$ is compared with -1.65 until it is less than -1.65 for the first time. Assuming that k ($2 \leq k < N$) cross-section elements fit the bill, the calculation criteria of the number of members k^* ($k^* \leq k$) in the core group are as follows:

$$k^* = \operatorname{argmax}_k \{t_{\hat{b}}(k)\} \quad \text{s.t.} \quad \min\{t_{\hat{b}}(k)\} > -1.65 \quad (14)$$

If $k^* = N$, the convergence club does not exist and the entire panel converges. When $k = 2$, the constraint condition $\min\{t_{\hat{b}}(k)\} > -1.65$ is not valid, then remove the highest ordered unit and repeat the above steps for the remaining units.

- (2) Club members: the cross-section elements outside the core group are added into the core group for log- t regression successively, and the value $t_{\hat{b}}$ is calculated. When it is greater than the critical value c (usually 0), the cross-section element is added into the convergence club.
- (3) Stop rule: after the first convergence club is formed, perform the log- t -test on the remaining units. If null Hypothesis H_0 is not rejected, the remaining units will be another convergence club. When the null hypothesis is rejected, repeat steps (1)–(3) for the remaining units.

2.2.3. Half-Life Cycle

The half-life cycle model is put forward by Art Schneiderman, and the main content is that the failure rate level presents a negative linear correlation in a certain time range. If the initial level defect rate and the lowest level defect rate is given, the time required to reduce the defect rate by half is half of the life cycle. The half-life cycle can be calculated by the rate of convergence, which means the number of days that a country reduces its income gap by half.

Y_t represents the defect action level at any time, γ is the steady value, Y_{t_0} shows the initial defect action level, and $t_{0.5}$ is the half-life cycle. Then the half-life cycle model can be described by the mathematical model:

After the first half-life cycle, the descending space is:

$$\ln Y_t - \gamma = \frac{1}{2} (\ln Y_{t_0} - \gamma) \quad (15)$$

After the number of half-life cycles i ($i = \frac{t-t_0}{t_{0.5}}, i \in R_0^+$), the remaining improvement space is:

$$\ln Y_t - \gamma = \left(\frac{1}{2}\right)^i (\ln Y_{t_0} - \gamma) \quad (16)$$

Steady-state calculation equation in the classic β convergence theory:

$$\ln(Y_{it}) - \ln(Y^*) = e^{-\beta(t-t_0)} (\ln(Y_{i_0}) - \ln(Y^*)) \quad (17)$$

By combining the half-life cycle and convergence rate, the equation can be obtained:

$$t_{0.5} = \frac{\ln(2)}{\beta} \quad (18)$$

Therefore, the time required to reduce the gap by half can be calculated based on the rate of convergence.

3. Data and Resources

To make the results more convincing, we need to use data from more provinces and more years. Figures are available only after Chongqing seceded from Sichuan province in 1997. Therefore, the data used in this paper are the per capita energy consumption (PEC) of China's 30 provinces from 1997 to 2016. The data are from the China energy statistical yearbook 1998–2017. Data for each year are obtained from the statistical yearbook of the following year. Because the data of some regions are not included in China's official statistics, Tibet, Hong Kong, Macao and Taiwan are not included in this study.

4. Results and Discussion

4.1. Identification of Outliers

The data for the time-series analysis in this paper are the PEC of 30 provinces. After the preliminary judgment of all the sequences, the Ningxia Hui Autonomous Region is taken as an example for analysis. The per capita energy consumption data of the Ningxia Hui Autonomous Region from 1995 to 2016 are shown in Table 1.

First, a sequence diagram is made according to the data as in Figure 1. It can be seen that there is an obvious abrupt change in 2003, which can be used to judge if there is an outlier based on further analysis.

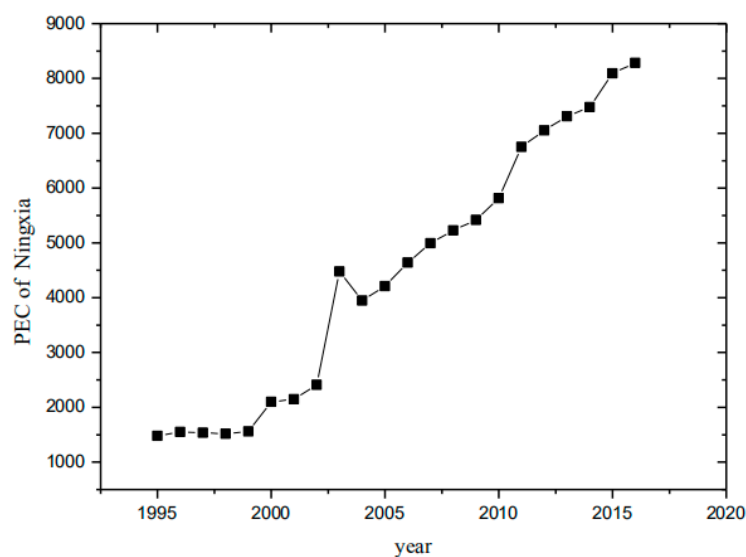


Figure 1. The per capita energy consumption data of Ningxia (kce/person).

Table 1. The per capita energy consumption data of the Ningxia Hui Autonomous Region (kce/person).

Year	Data	Year	Data	Year	Data
1995	1479.65	2003	4479.31	2011	6750.07
1996	1552.40	2004	3948.98	2012	7049.54
1997	1535.85	2005	4211.41	2013	7308.27
1998	1518.40	2006	4639.07	2014	7476.64
1999	1561.69	2007	4995.08	2015	8092.77
2000	2098.63	2008	5225.34	2016	8284.44
2001	2145.83	2009	5420.25		
2002	2409.09	2010	5815.69		

The identification process of outliers mainly adopts the joint estimation diagnosis method proposed by Chen and Liu (1993). The good statistical and anti-interference properties of the model can make it identify outliers under normal circumstances correctly. The first step is to establish the ARMA model by preliminarily judging if the data fit the AR (1) or MA (1) model based on the autocorrelation function and partial autocorrelation function. The result of establishing the ARMA model is shown in Table 2.

Table 2. The parameters of the autoregressive moving average (ARMA) model.

Case 1	Case 2
Model: 1 1 0	Model: 0 1 1
Coefficients:	Coefficients:
AR: 0.974419 ***	MA: 0.961220 *
AIC: −0.239257	AIC: 1.046726

Note: *** and * denote rejection of the null hypothesis at the 1% and 10% levels, respectively.

According to the Akaike Information Criterion (AIC), the AR (1) model is selected and it is a first-order autoregressive model. The parameters of the model are estimated as follows:

$$\hat{X}_t = 8.209311 + 0.974419x_{t-1}$$

According to the joint estimation method, the location and type of outlier are determined. The selection of critical value proposed by Chang (1988): $c = 4$ when the sensitivity is low, $c = 3.5$ when the sensitivity is medium and $c = 3$ when the sensitivity is high. In general, we name the critical value c as 3 when the sample size is less than 200.

Through the inspection, there is an outlier in Ningxia's per capita energy consumption in 2003, the Ningxia Hui Autonomous Region's τ value (3.9683) is greater than the critical value 3. Therefore, there is an outlier in Ningxia per capita energy consumption of 2003 and the type of it is an innovative outlier. After the identification of outliers is completed, the next section enters the data correction step.

4.2. Data Correction

In view of the data anomalies identified above, this paper adopts the regional convergence theory and then makes corrections according to the spatial correlation and geographical distribution characteristics between provinces and regions.

4.2.1. Club Grouping

The log- t method is used to divide the energy consumption per capita of 30 provinces into convergence clubs. Firstly, the per capita energy consumption of 30 provinces in 2016 is ranked from large to small. Ningxia ranked the highest. Taking Ningxia as the benchmark, we group Ningxia, Inner Mongolia, Qinghai, Xinjiang, Tianjin, Shanxi and Shanghai as the first core group through the core group test. Using the critical value $c = 0$, Liaoning, Hebei, Shandong, Jiangsu, Fujian, Shaanxi and Chongqing are added to the first club. The remaining provinces are gathered as the second group.

The division and order of clubs in 30 provinces are shown in Table 3. And the grouping of clubs is shown in Table 4.

Table 3. Division and order of clubs in 30 provinces.

Serial Number	Province	First Club		Second Club	
		Step 1	Step 2	Step 1	Step 2
1	Ningxia	benchmark	core		
2	Inner Mongolia	−0.1094	core		
3	Qinghai	−1.38054	core		
4	Xinjiang	−1.1309	core		
5	Tianjin	0.617951	core		
6	Shanxi	−0.17756	core		
7	Shanghai	2.484757	core		
8	Liaoning	1.562581	1.562581		
9	Hebei	1.270186	2.484757		
10	Shandong	1.848206	2.055578		
11	Jiangsu	1.786347	1.223246		
12	Zhejiang	0.96617	−1.00499	benchmark	
13	Heilongjiang	−0.57553	−6.50469	1.407396	
14	Beijing	−0.17778	−1.13035	6.157279	
15	Fujian	0.619555	1.902167		
16	Shaanxi	1.03126	0.447247		
17	Chongqing	1.394579	0.180409		
18	Jilin	0.763656	−3.69667	5.497718	
19	Guizhou	0.338821	−3.59548	7.292697	
20	Hubei	−0.00855	−2.50695	7.092816	
21	Guangdong	−0.72039	−5.46063	7.795714	
22	Gansu	−1.46008	−6.52179	7.023633	
23	Sichuan	−1.38843	−2.09056	7.041624	
24	Henan	−1.30784	−1.77379	7.538447	
25	Hunan	−0.79085	−0.53747	7.20602	
26	Yunnan	−1.3583	−5.71723	6.881611	
27	Hainan	−1.56274	−4.26987	6.73345	
28	Guangxi	−1.09245	−1.0436	6.628484	
29	Anhui	−1.95887	−8.91176	5.915468	
30	Jiangxi		−4.79133	5.766989	

The club grouping is shown in Figure 2.

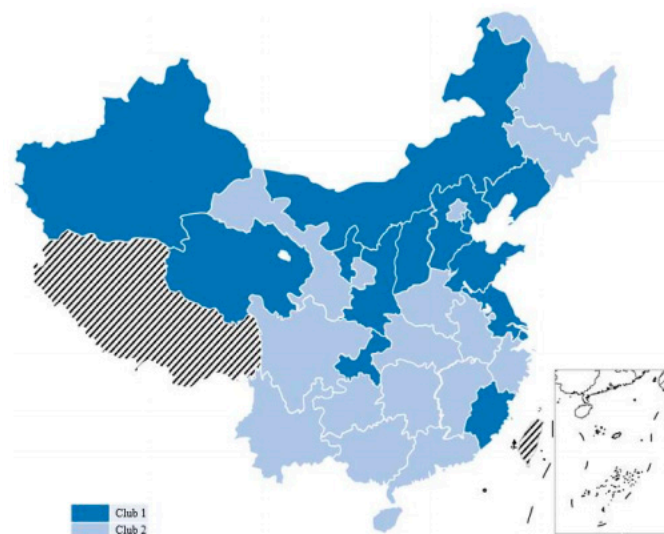


Figure 2. Results of the club grouping.

Table 4. Grouping of clubs.

Number	Members	The Number of Members	Category
Club 1	Ningxia, Inner Mongolia, Qinghai, Xinjiang, Tianjin, Shanxi, Shanghai, Liaoning, Hebei, Shandong, Jiangsu, Fujian, Shaanxi, Chongqing	14	High energy consumption
Club 2	Zhejiang, Heilongjiang, Beijing, Jilin, Guizhou, Hubei, Guangdong, Gansu, Sichuan, Henan, Hunan, Yunnan, Hainan, Guangxi, Anhui, Jiangxi	16	Low energy consumption

4.2.2. β Convergence Test

Based on the β convergence theory in 2.2.1, we conduct a regional convergence test within the divided clubs. The corresponding data are substituted into the equation for calculation, the results are shown in Table 5.

Table 5. Results of regional convergence test in clubs.

Test Parameters	Nationwide	First Club	Second Club
α	0.2211	0.2834	0.3011
t -Statistic	5.6626	4.4533	13.3257
Prob.	0.0000	0.0008	0.0000
β	0.0305	0.0450	0.0612
t -Statistic	3.0897	2.1954	5.7992
Prob.	0.0045	0.0485	0.0000
R^2	0.3849	0.5010	0.8957
Log likelihood	83.5074	40.1825	59.9100
F-statistic	17.5217	12.0503	120.2604
Prob (F-statistic)	0.0003	0.0046	0.0000
Durbin-Watson stat	1.2147	0.3032	1.0712

As the above results show, when the whole country is taken as the research object and clubs are not divided, the convergence rate of 30 provinces is 3.05%. When the clubs are grouped, the β value of the first club is 0.0450, so the convergence rate is 4.50%. The β value of the second club is 0.0612, so the convergence rate is 6.12%. Compared with the whole country, the convergence rate of two clubs was improved and the goodness of fit R^2 value was also improved.

4.2.3. Half-Life Cycle Correction

Based on the per capita energy consumption data of 30 provinces and cities in China from 1997 to 2016, the convergence rate of the whole and each club is calculated, and then the semi-life cycle is further calculated. The results are shown in Table 6.

Table 6. Results of the convergence rate and the semi-life cycle in clubs.

Club	Club 1	Club 2
β	4.50%	6.12%
γ	9.3628	8.3233
Half-life cycle (year)	15	11

The two clubs both have unconditional β convergence. The convergence rate of club 1 is 4.50% and the half-life cycle is 15 years, indicating that the gaps between different regions decrease at the convergence rate of 4.50% and the time required for the initial gap to be reduced to half size is 15 years. Similarly, for club 2, the convergence rate is 6.12%, and the half-life cycle is 11 years, indicating that

regional differences decrease at the convergence rate of 6.12% and it takes 11 years to reduce the initial gap by half.

The outlier identified in 4.1 is the data point of the Ningxia Hui Autonomous Region in 2003, which is in the first club with a convergence rate of 4.50%. The steady-state value calculated is 9.3628.

For the Ningxia Hui Autonomous Region, the per capita energy consumption in 2003 is revised to 2490.53 kce/person according to the half-life cycle equation with data from 1997. The data change condition after the correction is shown in Figure 3.

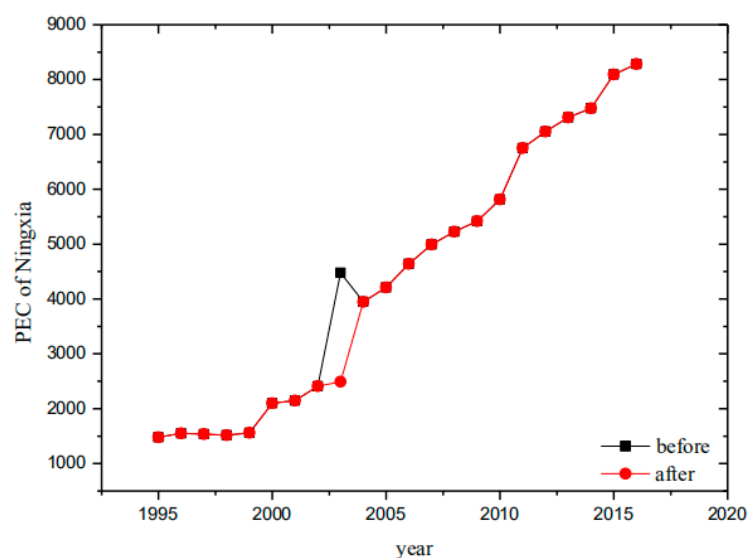


Figure 3. The per capita energy consumption data before and after correction of Ningxia (kce/person).

The data fluctuations become stable and no sharp changes occur.

The identification and correction process of outliers used in this study is a general system. For the identification process of the ARMA model, the larger the data volume is, the stronger its applicability and the better the data fitting effect will be. The theory of regional convergence is to divide the regions with similar initial conditions and structural characteristics into a group through geographical distribution characteristics and analyze the development status within the group which has good applicability.

5. Conclusions

Based on the theory of regional convergence in combination with a half-life cycle theory and time-series model of outlier's joint estimation method, this paper firstly identifies and corrects outliers of per capita energy consumption data of the Ningxia Hui Autonomous Region in 2003, then presents a reference energy outlier correction model according to regional convergence theory and characteristics of spatial distribution. The main conclusions of this paper are as follows:

- (1) For the time-series data fitting AR (1) model and through the outlier joint estimation diagnostic method, we calculate the τ value of Ningxia Hui Autonomous Region in 2003 and that is 3.97, which is greater than the critical value and identified as a mutational outlier.
- (2) β convergence exists nationwide with a convergence rate of 3.05%. According to the nonlinear time-varying factor model (log- t method), 30 provinces are divided into two convergence clubs. The convergence rate of the first club (high per capita energy consumption) is 4.5%, and that of the second club (low per capita energy consumption) is 6.12%. The convergence rate of the two clubs is higher than that of the whole country.
- (3) Based on the half-life cycle model and the convergence rate, the half-life cycle of the first club is 15 years and that of the second club is 11 years. By constructing the half-life cycle model

of β convergence theory, the revised data of the Ningxia Hui Autonomous Region represent 2490.53 kce/person.

- (4) The outliers identified in the paper are likely to be caused by human error, instrument failure and other errors in the process of data collection. This reminds us to attach importance to data collection, strengthen the supervision of data collection, and take measures such as multiple calculations to make the data real and effective.

This study takes China's per capita energy consumption as the research object, then identifies the outlier and corrects it based on regional convergence theory. Considering the characteristics of geographical distribution in the theory of regional convergence, the similarity within the club is used to correct the process, so as to provide a reference method for improving the stability of data. At the same time, the identification and correction process established in this paper is a set of general theories and has good applicability to other data. However, the correction model proposed in this paper is just a reference method for outliers based on regional convergence theory, indicating that it is not an absolute process. Meanwhile, the stability of specific correction results needs to be further studied.

Author Contributions: Conceptualization, N.L.; methodology, N.L., Y.L. and X.Z.; software, Y.L.; validation, Z.P.; formal analysis, Y.J.; data curation, N.L. and X.Z.; writing—original draft preparation, Y.L.; writing—review and editing, X.J., X.Z. and J.P.; supervision, H.M. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Acknowledgments: We gratefully acknowledge the financial support from the National Natural Science Foundation of China (No. 71603039, 5197602, 71828401). We are grateful to the editors and anonymous reviewers and to all those who have contributed to the improvement of this article.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

References

1. Solow, R.M. A Contribution to the Theory of Economic Growth. *Q. J. Econ.* **1956**, *70*, 65–94. [[CrossRef](#)]
2. Baumol, W.J. Productivity Growth, Convergence, and Welfare: What the Long-Run Data Show. *Am. Econ. Rev.* **1986**, *76*, 1072–1085.
3. Barro, R.J.; Sala-i-Martin, X. Convergence across States and Regions. *Brook. Papers Econ. Act.* **1991**, *1991*, 107–182. [[CrossRef](#)]
4. Roe, T. Determinants of Economic Growth: A Cross-Country Empirical Study. *Am. Political Sci. Rev.* **2003**, *92*, 145–477. [[CrossRef](#)]
5. Chambers, D.; Dhongde, S. Convergence in income distributions: Evidence from a panel of countries. *Econ. Model.* **2016**, *59*, 262–270. [[CrossRef](#)]
6. Iacovone, L.; Bayardo, L.F.S.; Sharma, S. *Regional Productivity Convergence in Peru*; Social Science Electronic Publishing: Washington, DC, USA, 2018.
7. List, J.A. Have air pollutant emissions converged among U.S. regions? Evidence from unit root tests. *South. Econ. J.* **1999**, *66*, 144–155. [[CrossRef](#)]
8. Mishra, V.; Smyth, R. Convergence in energy consumption per capita among ASEAN countries. *Energy Policy* **2014**, *73*, 180–185. [[CrossRef](#)]
9. Sheng, Y.; Shi, X.; Zhang, D. Economic growth, regional disparities and energy demand in China. *Energy Policy* **2014**, *71*, 31–39. [[CrossRef](#)]
10. Apergis, N.; Fontini, F.; Inchauspe, J. Integration of regional electricity markets in Australia: A price convergence assessment. *Energy Econ.* **2016**, *62*, 411–418. [[CrossRef](#)]
11. Solarin, S.A.; Gil-Alana, L.A.; Al-Mulali, U. Stochastic convergence of renewable energy consumption in OECD countries: A fractional integration approach. *Environ. Sci. Pollut. Res.* **2018**, *25*, 17289–17299. [[CrossRef](#)]
12. Galor, O. Convergence? Inferences from Theoretical Models. *Econ. J.* **1996**, *106*, 1056–1069. [[CrossRef](#)]
13. Sul, D. Transition Modeling and Econometric Convergence Tests. *Econometrica* **2007**, *75*, 1771–1855.

14. Phillips, P.; Sui, D. Economic Transition and Growth. *J. Appl. Econom.* **2010**, *24*, 1153–1185. [[CrossRef](#)]
15. Kim, Y.S. Electricity Consumption and Economic Development: Are Countries Converging to a Common Trend? *Energy Econ.* **2015**, *49*, 192–202. [[CrossRef](#)]
16. Parker, S.; Liddle, B. Economy-wide and manufacturing energy productivity transition paths and club convergence for OECD and non-OECD countries. *Energy Econ.* **2016**, *62*, 338–346. [[CrossRef](#)]
17. Mérida, A.L.; Carmona, M.; Congregado, E.; Golpe, A.A. Exploring the regional distribution of tourism and the extent to which there is convergence. *Tour. Manag.* **2016**, *57*, 225–233. [[CrossRef](#)]
18. Cuñado, J.; de Gracia, F.P. Real convergence in Africa in the second-half of the 20th century. *J. Econ. Bus.* **2006**, *58*, 153–167. [[CrossRef](#)]
19. Cuñado, J.; Gil-Alana, L.A.; de Gracia, F.P. Additional Empirical Evidence on Real Convergence: A Fractionally Integrated Approach. *J. Econ. Bus.* **2006**, *142*, 67–91.
20. Cuñado, J.; De Gracia, F.P. Real convergence in some Central and Eastern European countries. *Appl. Econ.* **2006**, *38*, 2433–2441. [[CrossRef](#)]
21. Romer, D. *Advanced Macroeconomics*, 4th ed.; McGraw-Hill Education: New York, NY, USA, 2001.
22. Seongman, M. Inter-Region Relative Price Convergence in Korea. *East Asian Econ. Rev.* **2017**, *21*, 123–146. [[CrossRef](#)]
23. Bergman, U.M.; Hansen, N.L.; Heeboll, C. Intranational Price Convergence and Price Stickiness: Evidence from Denmark. *Scand. J. Econ.* **2018**, *120*, 1229–1259. [[CrossRef](#)]
24. Fox, A.J. Outliers in Time Series. *J. R. Stat. Soc.* **1972**, *34*, 350–363. [[CrossRef](#)]
25. Tsay, R. Time Series Model Specification in the Presence of Outliers. *Publ. Am. Stat. Assoc.* **1986**, *81*, 132–141. [[CrossRef](#)]
26. Bruce, A.G.; Martin, R.D. Leave-k-out diagnostics for time series. *J. R. Stat. Soc.* **1989**, *51*, 363–424. [[CrossRef](#)]
27. Chen, C.; Liu, L.M. Joint Estimation of Model Parameters and Outlier Effects in Time Series. *J. Am. Stat. Assoc.* **1993**, *88*, 284–297.
28. Ping, C.; Jing, Y.; Li, L. Synthetic detection of change point and outliers in bilinear time series models. *Int. J. Syst. Sci.* **2015**, *46*, 284–293.
29. Yan, H.; Yang, B.; Yang, H. Outlier detection in time series data based on heteroscedastic Gaussian processes. *J. Comput. Appl.* **2018**, *76*. [[CrossRef](#)]
30. Sathe, S.; Aggarwal, C.C. Subspace histograms for outlier detection in linear time. *Knowl. Inf. Syst.* **2018**, *56*. [[CrossRef](#)]



© 2020 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/>).