Regional variability in water use intensity in China by clustering analysis
Xinghua Fan, Huihui Xu, Cheng Ning and Liangjie Wu

ABSTRACT

Water use intensity (WUI) reveals water withdrawals with respect to economic output. Decomposing WUI into factors provides inner-system information affecting the indicator. The present study investigates variability in WUI among provinces in China by clustering the principal components of the decomposed factors. Motivated by the index decomposition method, the authors decomposed WUI into seven factors: water use in agricultural, industrial, household and ecological sectors, exploitation rate of water resources, per capita water resources and population intensity. Those seven factors condense into four principal components under application of principal component analysis. Comprehensive WUI is calculated by these four components. Then the cluster analysis is applied to get different patterns in WUI. The principal components and the comprehensive intensity are taken as cluster variables. The number of clusters is determined to be three by applying the $k$-means clustering method and the $F$-statistic value. Variability in WUI is detected by implementing three clustering algorithms, namely $k$-means, fuzzy $c$-means and the Gaussian mixture model. WUI in China is clustered into three clusters by the $k$-means clustering method. Characteristics of each cluster are analyzed.

Key words | fuzzy $c$-means clustering, Gaussian mixture model, index decomposition, $k$-means clustering, principal component analysis, water use intensity

INTRODUCTION

Water shortage crisis has become a key issue restricting China’s sustainable development. In order to overcome the water crisis, China’s authorities advocated Water-Saving Society Construction in 1998. The government then issued total volume control and quotas for agriculture irrigation, for high water-consumption industries. Water management decisions must be made regarding the amount of water required to maintain diversity and ecological characteristics of the ecosystem. In 2012, the Chinese government proposed the development strategy of Construction of Water Ecological Civilization, which takes water ecological civilization construction as its nucleus. In the following year, the water use limiting index was set up and this target has been included in the future medium- and long-term plans for national economical and social development.

Water use has gained substantial attention. Focus lies mainly on the relationship between economic growth and water resources. A range of single or mixed methods, such as data envelopment analysis (Bian et al. 2014) and stochastic frontier analysis (Ferro et al. 2014), have been applied. Based on structural decomposition analysis, Cazcarro et al. (2015) found that growth in Spanish demand would have implied an increase in water. By using index number analysis, Luyanga et al. (2006) found that despite considerable increase in water tariffs, sectoral water efficiency in Namibia had not improved. Wang et al. (2014) analyzed the relationship between economic sectors and water use based on an input-output model. Some researchers have studied the factors of water use from the perspective of water footprint (Hubacek et al. 2009; Wang et al. 2015).
The goal in this study is to find a regional pattern of water use intensity (WUI) in China. As the largest developing country, there are great economic development disparities among provinces. Water resources exhibit highly heterogeneous geographical distributions. Most important of all, contradictions between environmental protection and economic development lead to variations in performance in different regions. Therefore, finding the differences in WUI among regions is vital for China’s water management policy so that it can adapt to different regions. Water use in this study refers to water diverted or withdrawn from its source. In this article, water use is equivalent to water supply assuming there is no loss of water. WUI is defined as the water use per economic output and thus it should be influenced by several factors. The goal is achieved through two steps. The first is to identify the contributions of different influencing factors and their combination to the WUI. The second is to apply cluster analysis to get different patterns in WUI.

The IPAT model will be extended to identify factors of WUI. The IPAT model (Chertow 2001) represents environmental impact (I) as the product of population (P), affluence (A) and technology (T). Commoner et al. (1971) were the first with the algebraic formulation \( I = PAT \) and the application of it to data analysis. Noting that not all factors are significant, principal component analysis (Syms 2008) will be applied to find out some new directions from those factors.

Cluster algorithm groups, or clusters, objects according to measured or perceived intrinsic characteristics or similarity. In recent years, a variety of clustering approaches (Berkhin 2006; Shamshirband et al. 2015) have been developed for applications in statistical data analysis. Three clustering approaches (k-means, fuzzy c-means and the Gaussian mixture model) are applied to determine regions with similar WUI patterns. Therefore, this study can fill the knowledge gaps of WUI in China.

**METHODOLOGY**

**WUI factor identification**

WUI measures the pressure of the economy on water resources in terms of the volume of water per unit of value added. Let \( I \) be the WUI. Then \( I = \frac{S}{Y} \), where \( S \) is water use in physical units and \( Y \) represents the value of gross domestic product (GDP).

Water is used both in social and economic activities and ecosystems. Ecological use of water is the water requirement for certain ecological functions, not for human activities. With the gradual deterioration of the ecological environment, ecological use of water becomes more and more important in maintaining diversity and the ecological characteristics of the ecosystem. The water use sector is broken down into four parts: agricultural, industrial, household and ecological fields. None of the four sectors interacts separately from all the others. Then \( S = \sum_{i=1}^{4} S_i \), where \( S_1, S_2, S_3 \) and \( S_4 \) are water use in agricultural, industrial, household and ecological sectors, respectively. WUI is related to water resources, the level of economic activity and population size. Motivated by the IPAT model, WUI can be decomposed as

\[
I = \frac{S}{Y} = \frac{\sum_{i=1}^{4} S_i}{Y} = \frac{\sum_{i=1}^{4} S_i}{G} \cdot \frac{G}{R} \cdot \frac{R}{P} \cdot \frac{P}{Y}
\]

where \( G \) signifies the water use of a region, \( R \) represents the total water resources, and \( P \) is the population size.

The formula gives a simple tool for understanding the components of water use intensity: the term \( \sum_{i=1}^{4} S_i/G \) is called the structure distribution of water use; \( S_i/G, i = 1, \ldots, 4 \) is agricultural, industrial, household and ecological structure of water use; \( G/R \) represents the exploitation rate of water resources; \( R/P \) means per capita water resources; and \( P/Y \) can be termed the population intensity of a region.

**Principal component analysis of WUI**

Let \( X \) be the data matrix of WUI whose \( i \)th row \( X_{(i)} \) represents water use factors of the \( i \)th province \((i = 1, \ldots, n)\), and the \( j \)th column \((j = 1, \ldots, p)\) gives a particular factor of water use. \( X \) is first centered and reduced. Mathematically, the principal components are defined by the eigenvectors of the covariance matrix of the water use data \( X'X \), where \( ' \) represents the transpose. Let \( W \) be the \( p \)-by\(-p \) matrix whose columns are the eigenvectors of \( X'X \). Then each row vector \( X_{(j)} \) of \( X \) is mapped to a new vector.
of principal component scores \( T_{(i)} = (t_1, \ldots, t_p)_{(i)} \) given by

\[
t_{(i)} = x_{(i)} \cdot w_{(k)}
\]

where \( w_{(k)} = (w_1, \ldots, w_p)_{(k)} \) (a \( p \)-dimensional vector called the weight or loading) is the \( k \)th eigenvector of \( XX \) (constrained to be a unit vector). The full principal components decomposition of \( X \) is given as

\[
T = XW
\]

Since the first few principal components contain more of the variance than the later components, \( x_{(i)} \) can be reasonably well approximated by including only the first few components \( t_{(i)} \) in (2).

**k-means clustering, fuzzy c-means clustering and the Gaussian mixture model**

In general, the goal of using cluster methods is to find an optimal grouping for which the observations or objects within each cluster are similar, but the clusters are dissimilar to each other.

Mathematically, \( k \)-means clustering is an optimization problem: find the \( k \)-cluster centers and assign the objects to the nearest cluster center, such that the squared error between the empirical mean of a cluster and the points in the cluster is minimized. In \( k \)-means clustering, clusters are represented by a central vector, which may not necessarily be a member of the data set.

Fuzzy \( c \)-means clustering allows one object to partially belong to more than one cluster (Sadri & Burn 2011). The aim of this clustering algorithm is to achieve a minimized total intra-cluster variance.

In Gaussian mixture models, clusters are defined as objects belonging most likely to the same Gaussian distribution (Krishnamurthy n.d.); \( k \)-clusters are found by learning the \( k \)-Gaussian distributions. In this paper, two-dimensional Gaussian mixture models are estimated for WUI.

**Parameter specification**

In this paper, the number of clusters for the \( k \)-means clustering is determined by the \( F \)-statistic value criteria. Suppose there are \( r \) clusters in the \( k \)-means clustering \((r = 2, \ldots, \lfloor n/2 \rfloor)\), where the operator \( \lfloor \cdot \rfloor \) means taking the integer part). For the \( k \)th cluster, let \( x^{(k)}_r \) be the \( r \)th object, \( n_k \) be the number of objects, and \( \bar{x}^{(k)} \) be its center. Let \( \bar{x} \) be the center of the whole data set. Then the \( F \)-statistic is defined as

\[
F(r) = \frac{\sum_{k=1}^{r} n_k \| \bar{x}^{(k)} - \bar{x} \|^2}{\sum_{k=1}^{r} \sum_{i=1}^{n_k} \| x_i^{(k)} - \bar{x}^{(k)} \|^2}
\]

The clustering is significant if \( F(r) > F_\alpha \), where \( \alpha \) is the significance level, \( F_\alpha \) is the upper critical value of the \( F \)-distribution with parameter \((r - 1, n - r - 1)\). The optimal number of clustering \( k \) is selected by maximizing \( F(r) - F_\alpha \).

**Experimental results**

**Data**

Seven influencing factors (agricultural structure of water use, industrial structure of water use, household structure of water use, ecological structure of water use, exploitation rate of water resources, per capita water resources and population intensity) were used to analyze WUI in China. Data were obtained from the China Statistical Yearbook for the time period 2004–2013, including 31 provinces except Hong Kong, Macao and Taiwan. Economic output is measured by current price and population size is counted by resident population.

All provinces shared a similar decreasing trend in WUI in the study period. The original data were averaged over the time period to decrease the temporal variability. Data then were normalized by Z-score methods.

**Principal component analysis**

The WUI indicators were computed at provincial level by using the normalized data. Let \( x_i^r \) be the normalized value of \( x_i \).

Eigenvalues of \( X'X \) are 2.8489, 1.4789, 1.1643, 0.7910, 0.5175, 0.1994 and 0. The values are relatively low after the fourth component. The sum of the first four eigenvalues exceeds 85\%, leading to the conclusion that the data set is
condensed into a reduced set of four new variables \( t_1, t_2, t_3, t_4 \), which are called principal components.

Component loadings are shown in Table 1. The columns refer to the principal components while the rows represent factors of WUI. The number marked with an asterisk indicates the greatest absolute value in each column.

By using Equation (3) the four principal components are given as

\[
\begin{align*}
    t_1 &= 0.5908x_1 - 0.6762x_2 - 0.797x_3 - 0.4784x_4 \\
    &\quad - 0.0968x_5' + 0.5028x_6' + 0.6640x_7' \\
    t_2 &= 0.3788x_1' - 0.697x_2' + 0.1918x_3' + 0.7985x_4' + 0.2153x_5' \\
    &\quad - 0.0457x_6' - 0.5562x_7' \\
    t_3 &= 0.0887x_1' + 0.1332x_2' - 0.4132x_3' - 0.1953x_4' \\
    &\quad + 0.8748x_5' - 0.3134x_6' - 0.2575x_7' \\
    t_4 &= -0.1066x_1' + 0.0964x_2' + 0.079x_3' - 0.0129x_4' \\
    &\quad + 0.2468x_5' + 0.8027x_6' - 0.2424x_7'.
\end{align*}
\]

Next, the authors found out the key factors in each principal component from the largest component loading. Component loadings stand for correlation coefficients of the principal components and each original variable. The greater the component loading, the more representative is the principal component as a factor of WUI. Table 1 reveals that the first principal component \( t_1 \) has the largest loading in the agricultural structure of water use. This suggests that the first component is primarily a measure of water use in the agricultural sector. Furthermore, the sum of the square of each column represents the variance of the component whose value is just the corresponding eigenvalue. The first principal component contributes most to \( 40.7\% \) of the total variance of WUI. Therefore, it is the dominant principal component and the authors named \( t_1 \) as the agriculture component of water use. A similar idea was used to identify other principal components. The authors regarded \( t_2 \) as the ecological component of water use. For \( t_3 \), the coefficient of \( x_5' \) is the largest number, which reflects the supply ability of water resources. The authors defined \( t_5 \) to be the water supply capacity component. For \( t_4 \), the coefficient of \( x_6' \) is much greater than the others, which indicates the abundance of water resources, so the authors called \( t_4 \) the carrying capacity component.

Comprehensive WUI can be obtained by

\[
F = \sum_{i=1}^{4} \frac{\lambda_i}{\sum_{j=1}^{4} \lambda_j} t_i
\]

or

\[
F = 0.4534t_1 + 0.2353t_2 + 0.1853t_3 + 0.1259t_4
\]

A negative value means the province has a lower WUI than average. The smaller the value, the better the performance in water use is. Furthermore, Equation (7) shows the relationship between the comprehensive WUI and its influencing factors. The coefficients of \( x_1', x_2', x_3', x_4' \) are positive while those of \( x_2', x_3', x_4' \) are negative, which means water use in the agricultural field, exploitation rate, per capita water resources and population intensity enlarge the WUI whereas water use in the industrial sector, households and the ecological field decrease it. Dominant in the positive factors and negative factors are agricultural structure and industrial structure of water use, respectively. Thus, provinces can lower their WUI mainly by optimizing their water use structure.

### Variability in WUI

Regional variability in WUI was studied by the clustering method. The four principal components \( t_1, t_2, t_3, t_4 \) and the

---

**Table 1 | Component loadings**

<table>
<thead>
<tr>
<th>Variable</th>
<th>Principal component</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1st</td>
</tr>
<tr>
<td>Agricultural structure of water use</td>
<td>0.9080*</td>
</tr>
<tr>
<td>Industrial structure of water use</td>
<td>-0.6762</td>
</tr>
<tr>
<td>Household structure of water use</td>
<td>-0.7970</td>
</tr>
<tr>
<td>Ecological structure of water use</td>
<td>-0.4784</td>
</tr>
<tr>
<td>Exploitation rate of water resources</td>
<td>-0.0968</td>
</tr>
<tr>
<td>Per capita water resources</td>
<td>0.5028</td>
</tr>
<tr>
<td>Population intensity</td>
<td>0.6640</td>
</tr>
</tbody>
</table>

* indicates the greatest absolute value in the column.
comprehensive component $I^*$ were selected as clustering variables. The clustering process was carried out on a PC using MATLAB 2012(a) software. Random initiation was used.

The clustering number was determined by using $k$-means clustering. For clustering numbers 2, ..., 5, the differences $F - F_a$ are 2.7319, 10.7960, 7.8990 and 5.4394. The optimal value $k_p = 3$ was obtained since it corresponded to the greatest difference.

Then $k$-means, fuzzy $c$-means and the Gaussian mixture model were applied to study the variability of WUI. Each clustering was repeated 50 times with clustering number $k_p = 3$. The one with the maximal $F$-statistic was selected as the clustering result and the $F$-statistic values averaged to provide the final estimate.

The average and the maximal $F$-statistic of the $k$-means, fuzzy $c$-means and Gaussian mixture model are 8.6609, 4.2398, 8.8242 and 12.0543, 8.3612, 10.3283, respectively. All clustering results are significant at the 95% confidence level. The greater the $F$-statistic, the better is the performance of the clustering algorithm, and $k$-means clustering and fuzzy $c$-means clustering performed better. But the maximal $F$-statistic was obtained in $k$-means clustering. Therefore, the final clustering result was determined by the $k$-means clustering with the maximal $F$-statistic.

Figure 1 shows the two-dimensional clustering result. The horizontal axis is the agricultural component while the vertical is the ecological component. The group has the best performance when it has the lowest WUI. Provinces in cluster 1 are located on the right side having the greatest agricultural component. This means they have the poorest quality in agricultural water use while provinces in cluster 3 have the best. The centers of the three clusters disperse in the agricultural component direction but are close in the ecological direction. This means it is mainly water use in the agricultural sector that makes the differences between clusters.

Figure 2 shows the schematic diagram of WUI. Areas with the same shaded pattern belong to the same cluster. One can see that members in the same cluster are mainly geographical neighbors, distributed from the northwest to the southeast according to the order of cluster 1 to cluster 3.

One can see members in each cluster in Figure 3. There are 5, 16 and 10 provinces in clusters 3, 2 and 1, respectively. Provinces in cluster 3 are relative economically developed areas in China while those in cluster 1 are comparatively undeveloped. Provinces in cluster 2 are developed or moderately developed areas.

Figure 3 shows comprehensive WUI for each province. The horizontal axis refers to provinces and the vertical axis represents comprehensive WUI. There is a downtrend in the WUI from cluster 1 to cluster 3. Each province in cluster 3 has negative comprehensive water use intensity. Thus provinces in cluster 3 have the best performance in WUI and those in cluster 1 have the worst.

From the above, one can see that provinces having the worst performance in WUI are located in the western and the northern parts of China. The best performing provinces are mainly developed areas (except Chongqing). But three developed provinces, namely, Guangdong, Jiangsu and Shandong are in the average-performing cluster.

**CONCLUSIONS AND DISCUSSION**

**Influencing variables of WUI**

Among the seven influencing variables, those in economic sectors or activities have more important roles. Equation (7) shows that agricultural structure of water use is the principal contributor to WUI, followed by those in industrial and household sectors.
Performance in WUI is almost proportional to the socio economic status of provinces. Developed areas in China have the best performance in WUI. Ten undeveloped provinces perform worst in WUI. WUI in other provinces is medium-sized.

Contributors to regional variations in WUI

Component loadings (Table 1) show that regional variations in WUI are mainly caused by the agricultural structure of water use and the ecological structure of water use.
ecological structure of water use plays a secondary but never negligible role.

Variability in WUI coincides with differences in ecological status. Neighboring provinces share similar WUI levels due to similar socio-economic and ecological status. WUI decreases from the northwest to the southeast, showing the characteristics of a highly heterogeneous geographical distribution. Having a vast territory, eco-environmental problems in China are closely related to regional geographical characteristics (Ouyang et al. 2000). Provinces in cluster 1 (except Tibet) suffer most from salinization and desertification. Many provinces in cluster 2 have serious soil erosion problems. The worse the eco-environmental problems, the less that water use has ecological functions and the greater is the WUI. Thus three developed provinces, Guangdong, Jiangsu and Shandong, unlike other developed provinces, are all in the average-performance cluster because eco-environmental problems in these provinces are moderate, neither very serious nor minor.

Implications for the water ecological civilization of China

The decomposition of the influencing factors of WUI and analysis of regional variability provides useful references for examining the effects of existing water-saving policies and supporting the formulation of future policies for the Water Ecological Civilization of China.

To lower water use intensity, the agricultural sector remains focal as the agricultural structure of water use has the largest and positive coefficient in comprehensive WUI. It is shown that China’s proportion of agriculture in GDP dropped to 9.2% in the year 2014, which is just a turning point of the national economy. Thus reducing the size of the agricultural sector, promoting water-saving irrigation and implementing water-saving agricultural technologies should be placed high on the agenda of China’s authorities.

The secondary activity to lower water use intensity should be industrial restructuring. More efforts should be paid to reducing or controlling the size of high water consumption industries such as the power, paper-making, metallurgy, chemicals and textile (silk) industries.

Since the ecological structure of water use plays a second important role in regional variability, local policy makers should take into consideration the construction of a water ecological civilization according to its own ecological characteristics. Although each province should implement the strictest water resources management system, there should be a greater difference in emphasis for different clusters. For example, provinces in cluster 1 should place more emphasis on restoration and protection of aquatic ecosystems. For cluster 2, it is water and soil conservation construction and construction of water conservancy areas for cluster 3.

Finally, it should be pointed out that a more detailed breakdown in WUI may provide more instruction for comparing WUI among those provinces. Also, other decomposition and clustering techniques could be applied to gain comparable results if compelling reasons are provided.

ACKNOWLEDGEMENTS

The research work is supported by the National Natural Science Foundation of China (51276081).

REFERENCES


Luyanga, S., Miller, R. & Stage, J. 2006 Index number analysis of Namibian water intensity. Ecological Economics 57 (3), 374–381.


First received 31 August 2015; accepted in revised form 22 February 2016. Available online 7 March 2016.