# Identification of key local factors influencing revenue water ratio of Korean cities using principal component analysis and clustering analysis

**S. Chung*, H. Lee*, M. Yu*, J. Koo*, I. Hyun** and H. Lee*****

*Department of Environmental Engineering, University of Seoul, 90, Jeonnong-dong, Dongdaemun-gu, Seoul, Korea (E-mail: *shinho@ene.uos.ac.kr; hkyoung-lee@hanmail.net; myong@uos.ac.kr; and jykoo@uos.ac.kr*)

**Department of Civil and Environmental Engineering, Dankook University, San-8, Hannam-dong, Yongsan-gu, Seoul, Korea (E-mail: *ihhyun@dankook.ac.kr*)

***Water Resources and Environmental Engineering Division, Korea Institute of Construction Technology, 2311, Daehwa-dong, Ilsan-gu, Goyang, Gyonggi-do, Korea (E-mail: *hdlee@kict.re.kr*)

**Abstract** In order to identify the relation between revenue water (RW) ratio and key local factors in a quantifiable way, 90 effect factors were considered as regional characteristics for 79 Korean cities. Seven statistically significant effect factors were chosen through correlation analysis. Three principal components independently influencing RW ratio were extracted by principal component analysis (PCA). The 79 cities were grouped into six clusters by *k*-means clustering (KMC) of the factor scores of the cities. Then key local factors were identified and their impacts were quantified by multiple regression analysis (MRA) and they were justified by *T*-test and *F*-test. The approach through correlation-PCA-KMC-MRA was proved to be one of scientific ways for identification of key local factors. According to the result, it was suggested that a shorter length of distribution system, a water supply with smaller number of bigger customer meters a and gravitational supply through reservoir would be advantageous from a RW ratio's point of view.

**Keywords** Key local factors; *k*-means clustering (KMC); multiple regression analysis (MRA); principal component analysis (PCA); revenue water; water loss management

## Introduction

Many researches related to water loss management are being conducted worldwide. A number of activities for reducing water losses and maximizing authorized billed consumption are on going in order to prevent huge economic loss. The ratio of non-revenue water was still 22.5% in Korea at the end of 2003 and it was equivalent to approximately $660 million when calculated on average tariff. Recently, inefficiency of percentage indicator for performance evaluation of distribution system was proved as it is dominated by consumption which is not a good explanatory factor, therefore gradually IWA's performance indicators, e.g. infrastructure leakage index (ILI) etc. are being applied among water utilities (Alegre *et al.*, 2000; Lambert *et al.*, 1999; Liemberger, 2002). However, in Korea as like many other countries, revenue water (RW) ratio is still being used for performance evaluation and it is not easy to shift this indicator to new one. Furthermore, percentage indicator may be used rationally but carefully if there is no dramatic change in consumption.

Korean water undertakers analyze their water distribution systems in many aspects. However, there has been no systematic research on why a certain RW ratio comes out from the system. All records kept in concerned departments are about accidentally occurring events which are not able to quantify.

Therefore, the relation between the RW ratio and regional characteristics should be investigated scientifically in the view of water loss reduction and the greatest influencing factors (key local factors) be identified by quantifiable ways. Then it would be possible

Water Science and Technology: Water Supply **Vol 5 No 3–4 pp 197–208** 

that more scientific water loss reduction strategy applicable to local situation is developed. Therefore, this study aimed to identify key local factors from reliable data and determine their impacts on RW ratio through statistical analyses for better water loss management.



## Method

### Target area and data

Statistical databases of the year 2000 showing regional characteristics for all the South Korean cities were used for this study. No city was excluded from this dataset. Since complete surveys of population, household, residence and territory for all the 79 cities had been conducted in year 2000, corresponding annual water statistics for the year 2000 were used here, despite the latest available water statistics for the cities was from 2002.

### Selection of effect factors

After simple calculation of the data obtained from the statistical database, 90 possible effect factors were qualitatively selected and they were classified into seven categories: scale of supply; density of water use; age of distributional district; residential type; residential level; construction and running cost of water supply; and others. Individual operating pressure was not included here, because the scope of target areas is city scale and not small sector scale. The effect factors are shown in Table 1. Linear relation of the effect factors on RW ratio were investigated by Pearson's correlation coefficients and they were tested at significance $\alpha = 0.05$. Seven statistically and logically meaningful effect factors were chosen.

### Principal component analysis and *k*-means clustering

The principal components analysis (PCA) which can reduce the dimensionality of a dataset consisting of a large number of interrelated variables while retaining little loss of information as least as possible (Jolliffe, 2002) was used for extracting principal components influencing RW ratio. The principal component axes were rotated by Varimax method which is one of the most efficient orthogonal rotation, three components having eigen value over 1 were selected as the principal components (Jolliffe, 2002; Kaiser, 1960). Efficiency of different clustering methods, i.e. agglomerative hierarchical clustering (AHC) and *k*-means clustering (KMC), using the factor scores of the components were compared graphically and the 79 cities were grouped into six city clusters.

### Multiple regression analysis of average properties of the city clusters

The key local factors influencing RW ratio were identified and their impacts were quantified by multiple regression analysis and they were justified by *T*-test and *F*-test. Several statistical values, i.e. coefficient of correlation ($r$), adjusted coefficient of determination (Adj $r$), root mean squared error (RMSE), mean absolute error (MAE) and variance inflation factor (VIF), were used for justification of this procedure. MRA results of other monothetic grouping methods, i.e. dividing into city groups with RW ratio such as $\leq$70%, 70–75%, 75–80%, 80–85%, 85–90% and $>$90% (comparison 1) and dividing into city groups with the same number of cities involved in the order of RW ratio (comparison 2), were compared. They were also compared with the MRA result conducted without city grouping (comparison 3).

**Table 1** Effect factors representing regional characteristics

| Symbol | Effect factors | Symbol | Effect factors |
|---|---|---|---|
| 1st category: scale of supply | | 4th category: residential type (continue) | |
| X1 | Total population | X46 | Rt. of condominium housing |
| X2 | Population served | X47 | Rt. of apartment housing |
| X3 | Supply area | X48 | Rt. of row housing |
| X4 | Total length of pipe line | X49 | Rt. of multi-h.hold housing |
| X5 | Length of dist. system | X50 | Rt. of the other housing |
| X6 | Length of dist. mains | 5th category: residential level | |
| 2nd category: density of water use | | X51 | Rt. of h.hold using standing kitchen |
| X7 | Population (recipient) density | X52 | Rt. of h.hold using conventional kitchen |
| X8 | Recipient per dist. system | X53 | Rt. of h.hold using w/o kitchen |
| X9 | Recipient per dist. mains | X54 | Rt. of h.hold using flush toilet |
| X10 | Production per dist. system | X55 | Rt. of h.hold using conventional toilet |
| X11 | Production per dist. mains | X56 | Rt. of h.hold living w/o toilet |
| X12 | Revenue water per dist. system | X57 | Rt. of h.hold using hot water bath facilities |
| X13 | Revenue water per dist. mains | X58 | Rt. of h.hold using cold water bath facilities |
| X14 | Large using building | X59 | Rt. of h.hold living w/o bath facilities |
| X15 | Estimated usage of large using building | X60 | Rt. of residence living $\leq$3 persons |
| X16 | Service connection density | X61 | Rt. of residence living 4–9 persons |
| 3rd category: age of distributional district | | X62 | Rt. of residence living $\geq$10 persons |
| X17 | Rt. of residence aged $\leq$10 years | X63 | Average residents number per residence |
| X18 | Rt. of residence aged 11–20 years | X64 | Rt. of residence with 1 room |
| X19 | Rt. of residence aged $\geq$21 years | X65 | Rt. of residence with 2–3 rooms |
| X20 | Average age of residence | X66 | Rt. of residence with 4–6 rooms |
| X21 | Rt. of comm. pipe aged $\leq$5 years | X67 | Rt. of residence with $\geq$7 rooms |
| X22 | Rt. of comm. pipe aged 6–10 years | X68 | Average room number per residence |
| X23 | Rt. of comm. pipe aged 11–15 years | 6th category: construction and running cost of supply | |
| X24 | Rt. of comm. pipe aged 16–20 years | X69 | Rt. of construction cost |
| X25 | Rt. of comm. pipe aged $\geq$21 years | X70 | Rt. of running cost |
| X26 | Average age of comm. pipe | X71 | Construction cost per total revenue |
| X27 | Rt. of dist. mains aged $\leq$5 years | X72 | Running cost per total revenue |
| X28 | Rt. of dist. mains aged 6–10 years | X73 | Construction cost per total length of pipe |
| X29 | Rt. of dist. mains aged 11–15 years | X74 | Running cost per total length of pipe |

**Table 1** (*continued*)

| Symbol | Effect factors | Symbol | Effect factors |
|---|---|---|---|
| X30 | Rt. of dist. mains aged 16 – 20 years | X75 | Construction cost per dist. system |
| X31 | Rt. of dist. mains aged ≥ 21 years | X76 | Running cost per dist. system |
| X32 | Average age of dist. mains | X77 | Construction cost per dist. mains |
| X33 | Rt. of dist. system aged ≤ 5 year | 7th category: others | |
| X34 | Rt. of dist. system aged 6 – 10 year | X78 | Running cost per dist. mains |
| X35 | Rt. of dist. system aged 11 – 15 year | X79 | Rt. of residential consumption |
| X36 | Rt. of dist. system aged 16 – 20 year | X80 | Rt. of institutional consumption |
| X37 | Rt. of dist. system aged ≥ 21 year | X81 | Rt. of commercial consumption |
| X38 | Average age of dist. system | X82 | Rt. of public bath consumption |
| X39 | Rt. of h.hold living in detached housing | X83 | Rt. of the other consumption |
| 4th category: residential type | | X84 | Rt. of low educational level |
| X40 | Rt. of h.hold living in condominium housing | X85 | Rt. of middle educational level |
| X41 | Rt. of h.hold living in apartment housing | X86 | Rt. of high educational level |
| X42 | Rt. of h.hold living in row housing | X87 | Rt. of 13 mm customer meter |
| X43 | Rt. of h.hold living in multi-h.hold housing | X88 | Average diameter of customer meter |
| X44 | Rt. of h.hold living in the other housing | X89 | Capacity of dist. reservoir per daily supply |
| X45 | Rt. of detached housing | X90 | Capacity of dist. reservoir per dist. system |

dist. = distribution; Rt. = Ratio; h.hold = household; condominium = apartment + row + multi household housing

## Result and discussion

### Relation between effect factors and revenue water ratio

Coefficient of correlations of 49 factors of the 90 effect factors appeared significant at 95% confidence level. No factor of the 1st category (scale of supply) showed significant as their coefficients were so small and no factor of the 6th category (construction and running cost of supply) showed its consistency. Recipients (residents served) per distribution system length (X8) was selected, which had a positive correlation as it represented density of water use. Ratio of residence aged 21 years and older (X19) having negative correlation was selected as representing age of distributional district and ratio of condominium housing (X46) having positive correlation was selected as representing residential type and ratio of household living without bath facilities (X59) having negative correlation was selected as representing residential level. Ratio of low educational level (X84) having a negative correlation was selected as related to resident's income level. Ratio of 13 mm customer meter (X87) having a negative correlation was selected as indicating potential leakage points, and the capacity of distribution reservoir per length of distribution system (X90) having positive correlation was selected as representing distribution type. Their relations are shown in Figures 1–7.

### Extraction of principal components

The above factors are not independent but interrelated to one another even though they indicate certain meaning of the relation she. Therefore, in order to extract new independent variables, so called, principal components which have maximum information of the original dataset, principal components analysis was conducted. The number of principal components having an eigenvalue over 1 was not enough without rotation.

Three principal components having explanatory variance ( * ) over 1 were extracted after Varimax rotation. The result of PCA was shown in Table 2. The 1st principal component (F1) represented city–rural strength, the 2nd (F2) represented distribution type and the 3rd (F3) represented the ratio of the small size customer meter. They were regarded as independently influencing factors to RW ratio. The greater the 1st and the 3rd principal components were, the lower RW ratio was, and vice versa. Similarly, the greater the 2nd principal component was, the higher the RW ratio was, and vice versa.

### Grouping city clusters by clustering analysis

In order to classify the cities having similar regional characteristics into a group, agglomerative hierarchical clustering (AHC) and *k*-means clustering (KMC) methods were used. Their clustering efficiencies were compared graphically after plotting the factor scores of

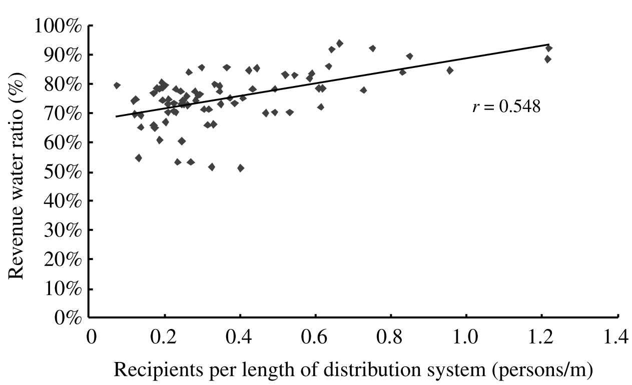


**Figure 1** Relationship between density of water use and revenue water ratio

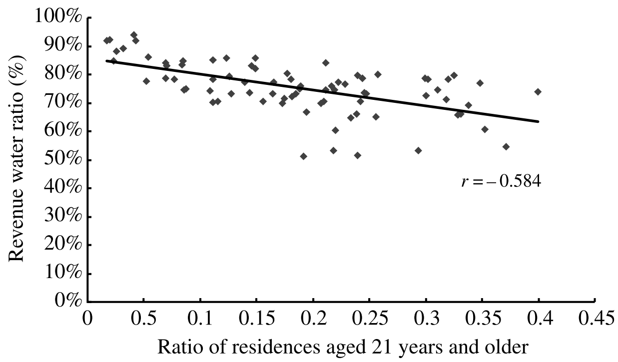


**Figure 2** Relationship between age of distributional district and revenue water ratio

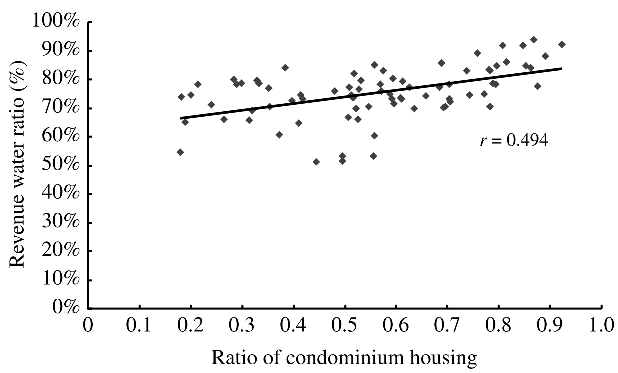


**Figure 3** Relationship between residential types and revenue water ratio

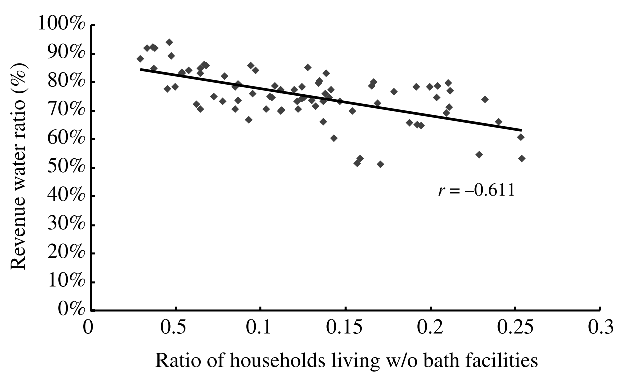


**Figure 4** Relationship between residential level and revenue water ratio

three principal components on three-dimensional space. AHC didn't make clusters efficiently during different number of clusters were tested in this study, but KMC did. Therefore, the cities were grouped by KMC and the number of clusters was set as 6 after comparing clustering efficiencies with different numbers of clusters. The comparison of KMC and AHC is shown in Figures 8 and 9. The 1st cluster represents the cities with light strength of rurality, the 2nd cluster represents old metropolitan and middle size regional cities and the 3rd cluster represents cities with high strength of rurality. The 4th cluster represents newly-developed cities dominated by condominium housing with relatively many small size customer meters, the 5th cluster represents newly-developed

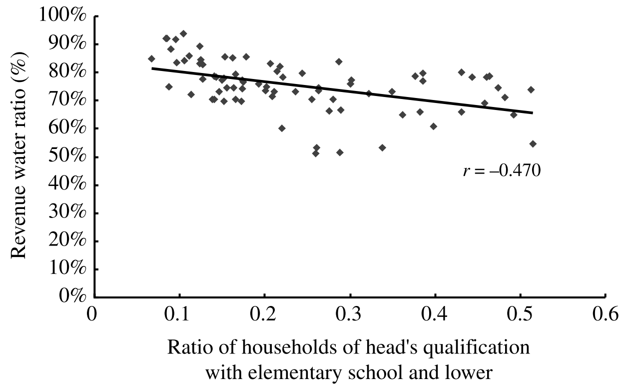


**Figure 5** Relationship between educational level and revenue water ratio

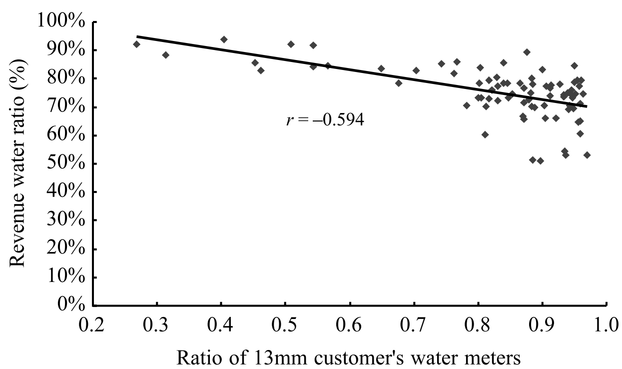


**Figure 6** Relationship between 13 mm customer meter and revenue water ratio

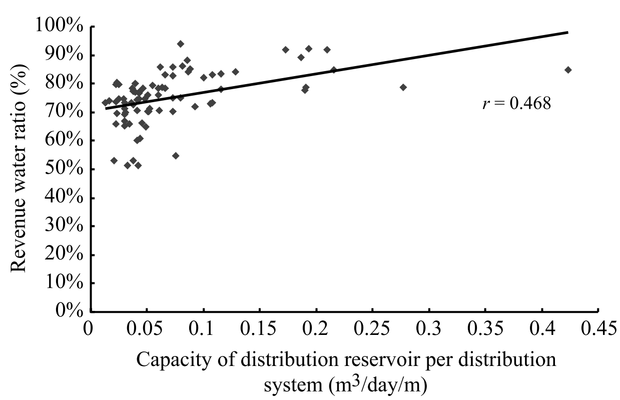


**Figure 7** Relationship between distribution types and revenue water ratio

cities dominated by condominium housing with a few small size customer meters, and the 6th cluster represents cities having extra ordinary capacity of distribution reservoir.

#### Multiple regression analysis of average properties of the six city clusters

The two variables were entered in multiple regression models because the increase of adjusted coefficient of determination was not big enough even if the number of variables increased to more than three. According to the result of MRA, two combinations of

**Table 2** Result of principal component analysis with/without rotation

| Effect factors | Factor loadings without rotation | | | | | | | Factor loadings after Varimax rotation | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | F1 | F2 | F3 | F4 | F5 | F6 | F7 | F1 | F2 | F3 | F4 | F5 | F6 | F7 |
| X8 | −0.839 | 0.380 | −0.140 | 0.344 | −0.117 | 0.023 | −0.015 | −0.378 | 0.428 | −0.346 | 0.739 | −0.083 | −0.022 | 0.006 |
| X19 | 0.961 | 0.122 | −0.118 | 0.015 | 0.017 | 0.213 | 0.047 | 0.825 | −0.309 | 0.266 | −0.244 | 0.152 | 0.264 | 0.000 |
| X46 | −0.943 | −0.201 | 0.128 | 0.089 | 0.148 | 0.012 | 0.155 | −0.888 | 0.239 | −0.249 | 0.245 | −0.026 | 0.018 | 0.177 |
| X59 | 0.917 | 0.197 | −0.042 | 0.178 | 0.286 | −0.065 | −0.032 | 0.754 | −0.233 | 0.306 | −0.173 | 0.502 | 0.028 | −0.001 |
| X84 | 0.909 | 0.313 | −0.175 | −0.040 | −0.128 | −0.110 | 0.123 | 0.939 | −0.172 | 0.203 | −0.164 | 0.050 | −0.066 | 0.119 |
| X87 | 0.700 | −0.015 | 0.695 | 0.132 | −0.093 | −0.001 | 0.006 | 0.292 | −0.121 | 0.924 | −0.197 | 0.079 | 0.017 | −0.005 |
| X90 | −0.675 | 0.655 | 0.255 | −0.206 | 0.083 | 0.022 | −0.002 | −0.254 | 0.934 | −0.113 | 0.214 | −0.067 | −0.017 | 0.005 |
| Eigenvalue | 5.128 | 0.766 | 0.631 | 0.220 | 0.150 | 0.063 | 0.043 | 3.213* | 1.307* | 1.255* | 0.807 | 0.296 | 0.076 | 0.045 |
| % variance | 73.3 | 10.9 | 9.0 | 3.1 | 2.1 | 0.9 | 0.6 | 45.9 | 18.7 | 17.9 | 11.5 | 4.2 | 1.1 | 0.6 |
| Cumul. % | 73.3 | 84.2 | 93.2 | 96.4 | 98.5 | 99.4 | 100.0 | 45.9 | 64.6 | 82.5 | 94.0 | 98.3 | 99.4 | 100.0 |

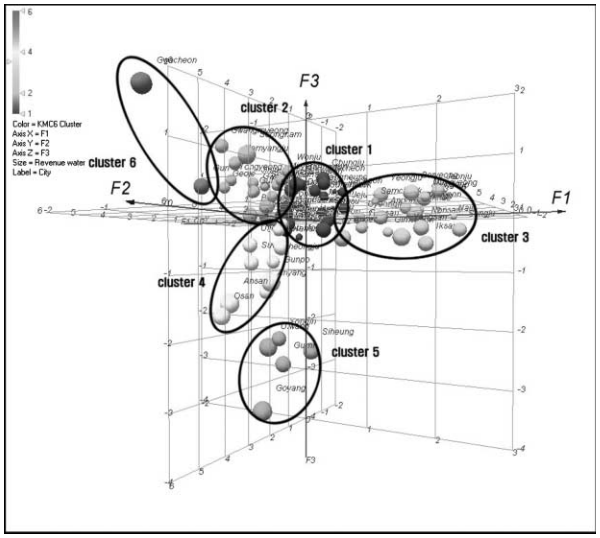


**Figure 8** Three-dimensional scatter plot of the factor scores by KMC (6 clusters)

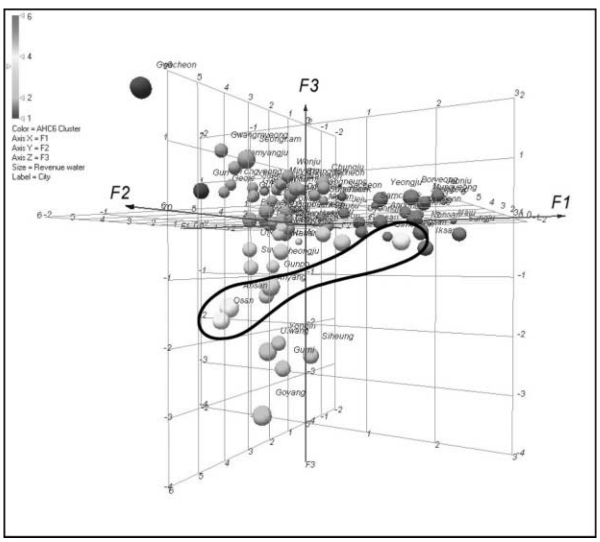


**Figure 9** Three-dimensional scatter plot of the factor scores by AHC (6 clusters)

recipients per distribution system length (X8) with ratio of 13 mm customer meter (X87), and ratio of 13 mm customer meter (X87) with capacity of distribution reservoir per length of distribution system (X90) showed the best fit for all statistics when the cities were classified by KMC. On the other hand, the results of comparison 1 and 2 didn't show good fit as they failed in *t*-test and/or VIF even with all other combination of variables. The coefficients of correlation and determination of comparison 3 were so low when compared to previous ones. Therefore, it could be thought that the three effect factors (recipient per distribution system length (X8), ratio of 13 mm customer meter (X87) and capacity of distribution reservoir per length of distribution system (X90)) are the key local factors influencing RW ratio. The results of MRA by KMC and three comparisons were given in Table 3.

**Table 3** Comparison of multiple regression analysis results

| Method | KMC | | Comp.1 | Comp.2 | Comp.3 |
|---|---|---|---|---|---|
| Variables | X8, X87 | X87, X90 | X46, X87 | X59, X90 | X59, X87 |
| $r$ | 0.980 | 0.973 | 0.991 | 0.934 | 0.673 |
| Adj. $r^2$ | 0.934 | 0.912 | 0.969 | 0.787 | 0.438 |
| RMSE | 0.018 | 0.021 | 0.019 | 0.043 | 0.069 |
| MAE(%) | 1.4% | 1.5% | 1.5% | 3.8% | 7.5% |
| *F*-test | OK | OK | OK | OK | OK |
| *T*-test | OK for all | OK for all | Failed for 1 | Failed for all | OK for all |
| VIF | OK for all | OK for all | Failed for all | OK for all | OK for all |

**Table 4** Multiple regression models for RW ratio by KMC

| Adj. $r^2$ | Model |
|---|---|
| 0.934 | $Y = 0.18751X8 - 0.17191X87 + 0.82528$ |
| 0.912 | $Y = -0.27901X87 + 0.31140X90 + 0.97026$ |

With the two variables selected above, multiple regression models for RW ratio were composed as shown in Table 4.

Observed RW ratio and predicted values by KMC are plotted in Figure 10 and show a good fit. In contrast, the equivalent data without clustering are plotted in Figure 11 and show poor fit.

According to these models by KMC, X8 and X90 give a positive impact on RW ratio (*Y*), while X87 gives a negative impact. In order to achieve high RW ratio, recipient per distribution system length (X8) and capacity of distribution reservoir per length of distribution system (X90) should be greater and the ratio of small size customer meter (X87) should be smaller. This means it is better to make the length of the distribution system shorter and the capacity of the distribution reservoir greater in order to make X8 and X90 high. But recipient population cannot be adjusted by force although it can be controlled at the planning stage of development of the area. Similarly, a small number of big size

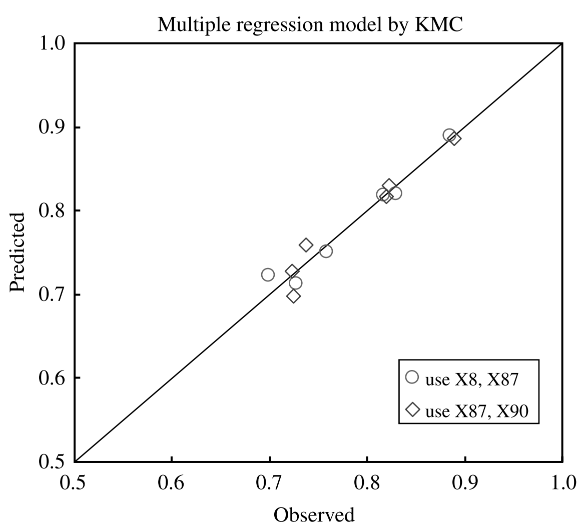


**Figure 10** Comparison of observed and predicted values of RW ratio by KMC

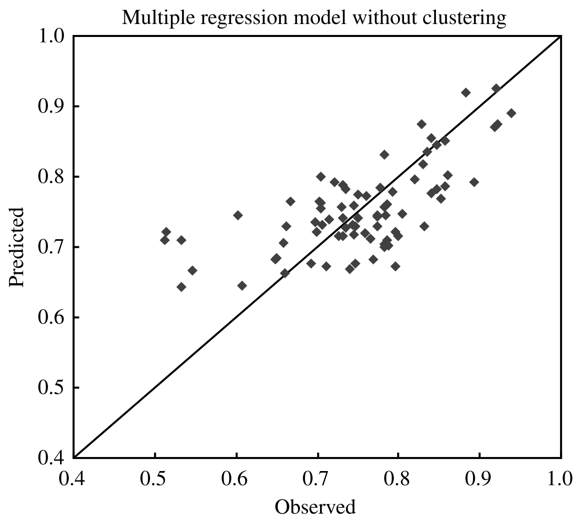


**Figure 11** Comparison of observed and predicted values of RW ratio without clustering (comp. 3)

customer meters would be better compared to many small size customer meters. A high ratio of small size customer meter indicates that there are more connections in pipes and implies more potential of leakage (WRc, 1994). It is common sense that service connection contributes water loss much more than distribution mains (Lambert *et al.*, 1999). However, it must not be thought that the ratio of small size customer meter can be lowered by replacement of small size customer meter a large one because leakage potential from connection points would not be reduced.

## Discussion

Lambert *et al.* (1999) and Lambert and Hirner (2000) have identified key local factors influencing real losses, which are continuity of supply, length of mains, number of service connections, location of customer meter on service connection and average operating pressure. In this study, similar key local factors were identified by a different approach. Continuity is not worthy of consideration in Korea because already all the service area is being continuously supplied. Length of distribution system (sum of the length of distribution main and communication pipe) instead of distribution mains was identified here. Capacity of distribution reservoir is thought to be related to stability of pressure rather than the pressure itself. Although water is distributed through reservoir, high pressure may be experienced. However, fluctuation of pressure in the case of supplying through the reservoir is much less than that in the, case of direct pumping. Also, the ratio of small size customer meter was identified instead of the number of service connections. Although service connection density expressed as number of service connections per distribution main (X16) was included on the list of 90 effect factors, it was not selected because of a statistically weak relation. It was thought that the service connection density cannot differentiate an area having a certain number of big size customer meters mostly from other areas having a similar number of small size customer meters mostly with similar distribution main, although the scale of supply would be much different each other. In that case, real losses per service connection would be similar but the RW ratio would be different. Therefore, it may be thought that the ratio of small size customer meters is more related to the RW ratio while service connection density is more related to the amount of real losses. Key local factors extracted in this study, however, may not be adjusted easily at the time of operation but can be considered at the time of planning.

## Conclusions

The approach using correlation, principal component analysis, *k*-means clustering followed by multiple regression analysis was proved to be one of the scientific ways for identifying key local factors influencing revenue water ratio.

Density of water use expressed as recipients per distribution system length, potential of leakage points expressed as a ratio of 13 mm customer meter and distribution type expressed as capacity of distribution reservoir per distribution system length were identified as key local factors. And RW ratio was found to be influenced by such less-flexible local factors.

According to the series of analysis, shortening the length of distribution system, supplying the same amount of water with a small number of bigger customer meters rather than the large number of small customer meters and supplying water by gravity through reservoir rather than direct pumping would be beneficial for water loss management.

## Acknowledgements

This research was supported by a grant (4-2-2) from Sustainable Water Resources Research Center of 21st Century Frontier Research Program.

## References

Alegre, H., Hirner, W., Baptista, J.M. and Parena, R. (2000). *Performance Indicators for Water Supply Services*, IWA publishing ‘Manuals of best practice’ series, IWA, London.

Jolliffe, I.T. (2002). *Principal component analysis*, Springer series in statistics, Springer-Verlag, New York.

Kaiser, H.F. (1960). The application of electronic computers to factor analysis. *Educ. Psychol. Meas.*, **20**, 141–151.

Lambert, A.O. and Hirner, W.H. (2000). *Losses from Water Supply System: Standard Terminology and Performance Measure*, IWA Blue Pages, IWA, London, UK.

Lambert, A.O., Brown, T.G., Takizawa, M. and Weimer, D. (1999). A review of performance indicators for real losses from water supply systems. *Journal of Water SRT-Aqua*, **48**, 227–237.

Liemberger, R. (2002). Do you know how misleading the use of wrong performance indicators can be?, Presented at IWA Managing Leakage Conference, Cyprus.

WRC (1994). *Managing leakage series*, Engineering and operations committee, Water Research Centre.