Yeung, Daisy & Skitmore, Martin
(2012)
A method for systematically pooling data in very early stage construction price forecasting.
Construction Management and Economics, *30*(11), pp. 929-939.

A METHOD FOR SYSTEMATICALLY POOLING DATA IN VERY EARLY-STAGE CONSTRUCTION PRICE FORECASTING

Daisy K. L. Yeung[1] and Martin Skitmore[2]

[1]Department of Building and Construction, City University of Hong Kong, Hong Kong.

[2]Professor of Construction Economics and Management, School of Civil Engineering and Built Environment, Queensland University of Technology, Brisbane, Queensland 4001, Australia

30 August 2012

A METHOD FOR SYSTEMATICALLY POOLING DATA IN VERY EARLY-STAGE CONSTRUCTION PRICE FORECASTING

ABSTRACT

Client/owners usually need an estimate or forecast of their likely building costs in advance of detailed design in order to confirm the financial feasibility of their projects. Because of their timing in the project life-cycle, these early stage forecasts are characterized by the minimal amount of information available concerning the new (target) project to the point that often only its size and type is known. One approach is to use the mean contract sum of a sample, or base-group, of previous projects of a similar type and size as the estimate needed. Bernoulli's law of large numbers implies that this base group should be as large as possible. However, increasing the size of the base group inevitably involves including projects that are less and less similar to the target project. Deciding on the optimal number of base group projects is known as the homogeneity or pooling problem.

A method of solving the homogeneity problem is described involving the use of closed from equations to compare three different sampling arrangements of previous projects for their simulated forecasting ability by a cross validation method, where a series of targets are extracted, with replacement, from the groups and compared with the mean value of the projects in the base-groups. The procedure is then demonstrated with 450 Hong Kong projects (with different project types: Residential, Commercial centre, Carparking, Social community, School, Office, Hotel, Industrial, University and Hospital) clustered into base-groups according to their type and size.

Keywords: cross validation, data pooling, early stage estimating, homogeneity, closed form equations.

INTRODUCTION

As Flanagan (1980) points out, "a reliable price prediction for a proposed building is probably one of the most important tasks ... because a client will often base his investment decision on this forecast". However, construction cost is notoriously difficult to forecast, especially in the early stage of projects, as most of the information concerning the new (target) project is very scarce (Skitmore, 1991).

One approach is to use the mean contract sum of a group of similar projects, or base group, to the target project. As introduced by Skitmore (2001), Beeston (1974) has urged the use of as large a base group as possible for analysis in order to reduce the effects of sampling bias. However, as originally pointed out by Flanagan (1980), this produces a paradoxical situation sometimes termed the homogeneity or pooling problem. Ideally, the forecaster would use a base group comprising a sample of similar projects to the target project, ie., of similar functional and technological type, size, geographical location, etc. The assumption is that the closer the characteristics of the base group match the target project, the better the ensuing forecast will be. However, the closer the base group is made to match the target project, the smaller the base group becomes, and the greater becomes the sampling bias involved. Clearly, the solution to this dilemma is to somehow trade-off the biases created by using too small a base group with the biases created by using an unrepresentative sample.

Since Flanagan, Skitmore (2001) offered an approach to solving the problem in the risk analysis context by empirically examining the effects of all possible pooling combinations on forecasting errors with a view to selecting the data pooling arrangement that best minimises the spread of errors. To do this, he divided the base group data into five groups (1) construction floor area, (2) contract sum, (3) nature of works, (4) project type, and (5) number of bidders. Then, a cross validation (leave-one-out) method was used to identify the best pooling arrangement from a variety of combinations of data pooling arrangements. Later preliminary work by Yeung and Skitmore (2005) extended this to an example involving just three groups: (1) construction floor area, (2) building type, and (3) client type. For both studies, however, the cross validation method was to be applied manually – a very laborious and time-consuming activity that places severe limits on the amount of practicable analysis that can be undertaken.

A solution is provided in the form of a set of closed form equations, which enables the analyses to be conducted on any scale desired. As the mathematical equivalent of previous work, this still provides the results of leaving one project out, comparing the mean value of the remaining projects to the one left out, and then repeating the process with replacement. Carrying out this procedure for different base group compositions, then enables the user to (1) assess the performance of forecasted construction cost for different types of cost data groupings, and (2) identify the cost data groupings that provide the best forecasting results. In short, the situation is considered where a very early stage construction price forecast is needed and where the mean of a set of prices of similar projects is used for the forecast. Hence, the method identifies the projects that constitute the set of similar projects to use. Later, the method is demonstrated in the detailed analysis of winning bid prices of 450 Hong Kong building projects.

For clarity, the following terminology is used:

Construction cost - the amount paid by the client/owner to the constructor and is used synonymously here with the term *price*.

Forecast - used synonymously with the term *pretender estimate* of the project price

Target – the new project for which a cost forecast is needed

Base - a single historical project for which the cost is already known

Base-group – a set of base projects

Pooling method - the strategy used for determining the projects that comprise the base-group

FORMULAE

In general terms, Skitmore's (2001) analysis reduces to the consideration of two base-groups. One base-group contains projects of the same characteristics as the target project, while the other base-group contains projects that have (possibly only marginally) different characteristics to the target project. Let X and Y be independent random variables denoting the known construction cost of the projects in the first and second base-group respectively. Let these individual project costs be denoted by observations $x_1, x_2, \ldots, x_n$ and $y_1, y_2, \ldots, y_m$

respectively. A forecast is therefore needed of the future, as yet unknown observation x_{n+1}. Using the mean of the recorded observations as the forecast, three types of forecast are considered:

(I) $\hat{x}_{n+1} = \bar{x}$

(II) $\hat{x}_{n+1} = \bar{y}$

(III) $\hat{x}_{n+1} = \dfrac{n\bar{x} + m\bar{y}}{m+n}$

where (I) denotes the situation where the forecast is the mean value of the target group of projects, (II) the mean value of the non-target project and (III) the mean value of a mixture of target and non-target projects. Assuming goodness-of-fit is measured by the square of the error of forecast, i.e., $\left(x_{n+1} - \hat{x}_{n+1}\right)^2$, we seek to select the type of forecast that minimizes this error of forecast. Using the leave-one-out or cross validation method provides a simulated error, by placing one project in a hold-out sample, using the remaining projects to generate a forecast, applying the forecast to the hold-out sample and recording the error. The hold-out project is then returned to the database and replaced by another project from the database. The process (Figure 1) is then repeated until all the projects in the database have been used – the errors obtained in this way being a measure of the simulated forecasting ability of the forecasting method used.

Type I

For (I), the simulated forecast error is simply given by the mean square

$$msq_{(I)} = \frac{1}{n}\sum_{i=1}^{n}\left(x_i - \bar{x}_i\right)^2 \tag{1}$$

where $\bar{x}_i$ denotes the mean of the $x_1, x_2, \ldots, x_n$ observations excluding the ith observation. That is

$$\begin{aligned} msq_{(I)} &= \frac{1}{n}\sum_{i=1}^{n}\left(x_i - \frac{n\bar{x} - x_i}{n-1}\right)^2 \\ &\Rightarrow \frac{n}{n-1} S_X^2 \end{aligned} \tag{2}$$

where $S_X^2 = \sum_{i=1}^{n}\left(\dfrac{\bar{x} - x_i}{n-1}\right)^2$.

Type II

For II, the equivalent mean square is given by

$$\begin{aligned} msq_{(II)0} &= \frac{1}{n}\sum_{i=1}^{n}\left(x_i - \bar{y}\right)^2 \\ &\Rightarrow \left(\bar{x} - \bar{y}\right)^2 + \frac{n-1}{n} S_X^2 \end{aligned} \tag{3}$$

However, for a complete simulation of $\hat{x}_{n+1}$ it is necessary to consider all possible combinations of unknown $y_1, y_2, \ldots, y_m$ observations, e.g., for one missing observation

$$msq_{(II)1} = \frac{1}{mn}\sum_{i=1}^{n}\sum_{j=1}^{m}\left(x_i - \bar{y}_j\right)^2 \tag{4}$$

where $\bar{y}_j$ denotes the mean of the $y_1, y_2, \ldots, y_m$ observations excluding the *j*th observation. That is

$$\begin{aligned} msq_{(II)1} &= \frac{1}{mn}\sum_{i=1}^{n}\sum_{j=1}^{m}\left(x_i - \frac{m\bar{y} - y_j}{m-1}\right)^2 \\ &\Rightarrow \left(\bar{x} - \bar{y}\right)^2 + \frac{n-1}{n}S_X^2 + \frac{1}{m(m-1)}S_Y^2 \end{aligned} \tag{5}$$

Likewise, for two missing observations

$$\begin{aligned} msq_{(II)2} &= \frac{2}{m(m-1)n}\sum_{i=1}^{n}\sum_{j1=1}^{m}\sum_{\substack{j2=1 \\ j2\neq j1}}^{m}\left(x_i - \frac{m\bar{y} - y_{j1} - y_{j2}}{m-2}\right)^2 \\ &\Rightarrow \left(\bar{x} - \bar{y}\right)^2 + \frac{n-1}{n}S_X^2 + \frac{2}{m(m-2)}S_Y^2 \end{aligned} \tag{6}$$

which generalises, for *p* missing observations, to

$$msq_{(II)p} = \left(\bar{x} - \bar{y}\right)^2 + \frac{n-1}{n}S_X^2 + \frac{p}{m(m-p)}S_Y^2 \tag{7}$$

giving, for all combinations of missing observations

$$ssq_{(II).} = \sum_{p=0}^{m-1} q_p\left(\left(\bar{x} - \bar{y}\right)^2 + \frac{n-1}{n}S_X^2 + \frac{p}{m(m-p)}S_Y^2\right) \tag{8}$$

where $q_p = \dfrac{nm!}{p!(m-p)!}$ and therefore $msq_{(II).} = \dfrac{ssq_{(II).}}{\sum_{p=0}^{m-1} q_p}$. Note that to avoid computational overflow it is preferable to use $q_p = \exp\left(\ln n + \sum_{i=1}^{m}\ln i - \sum_{i=0}^{p}\ln i - \sum_{i=0}^{m-p}\ln i\right)$, where $\ln 0 = \ln 1$

Type III

For III,

$$msq_{(III)0} = \frac{1}{n}\sum_{i=1}^{n}\left(x_i - \frac{n\bar{x} + m\bar{y} - x_i}{m+n-1} \right)^2 \Rightarrow \frac{1}{(m+n-1)^2}\left[m^2(\bar{x}-\bar{y})^2 + (m+n)^2\left(\frac{n-1}{n}\right)S_X^2 \right] \quad (9)$$

with

$$msq_{(III)1} = \frac{1}{mn}\sum_{i=1}^{n}\sum_{j=1}^{m}\left(x_i - \frac{n\bar{x} + m\bar{y} - x_i - y_j}{m+n-2} \right)^2 \Rightarrow \frac{1}{(m+n-2)^2}\left[(m-1)^2(\bar{x}-\bar{y})^2 + (m+n-1)^2\left(\frac{n-1}{n}\right)S_X^2 + \left(\frac{m-1}{m}\right)S_Y^2 \right] \quad (10)$$

which generalises, for *p* missing observations, to

$$msq_{(III)p} = \frac{1}{(m+n-p-1)^2}\left[(m-p)^2(\bar{x}-\bar{y})^2 + (m+n-p)^2\left(\frac{n-1}{n}\right)S_X^2 + p\left(\frac{m-p}{m}\right)S_Y^2 \right] \quad (11)$$

giving, for all combinations of missing observations

$$ssq_{(III).} = \sum_{p=0}^{m-1} q_p\left(msq_{(III)p}\right) \quad (12)$$

where $q_p = \frac{nm!}{p!(m-p)!}$ as before, and $msq_{(III).} = \frac{ssq_{(III).}}{\sum_{p=0}^{m-1} q_p}$.

PROCEDURE

Data sampling

To demonstrate the procedure involved, the winning bid prices are analysed for 450 Hong Kong building projects collected from the: (1) Private Sector (one of the largest cost consultants firms in Hong Kong) and (2) Public Sector (government department of the Hong Kong Special Administrative Region of the People's Republic of China). The projects range from year 1995 to 2006 and comprise information regarding (1) project type, (2) preliminary project specification, (3) construction floor area (CFA), (4) bid date and (5) bid price per HK$/1000m^2 adjusted to the 3rd quarter of 2006.

Base-group clustering

The collected data were clustered hierarchically (after Berkhin 2006) into two levels with (1) project type clusters and (2) specification type and CFA type clusters. The clustering details are summarised in Table 1.

For the 450 projects, a total of 10539 base-groups were generated as detailed below:

Level 1: Project type grouping

The 450 projects were grouped into ten different building types of

1. Residential [R]
2. Commercial centers [C]
3. Car parking [P]
4. Social community centers [L]
5. Schools [S]
6. Offices [O]
7. Hotels [T]
8. Industrial [I]
9. Universities [U] and
10. Hospitals [H].

The base-groups were then identified for each target project type, comprising: (1) the target project type itself, (2) all combinations of project types excluding the target type and (3) the target type and all combinations of all the other project types. For example, when the target [TG] is a Residential project, the base-groups comprise: (1) the project type R group (1 base-group), (2) the type C group, the type P group, ... , the type H group, the pooled type [C] and P groups, the pooled type C and L groups, ..., the pooled type C and H groups, the pooled type P and L groups, ..., the pooled type C, P and L groups, etc (511 base-groups) and (3) the pooled type R and C groups, the pooled type R and P groups, ..., the pooled type R and H groups, the pooled type R, C and P groups, the pooled type R , C and L groups, ... , the pooled type R, C and H groups, the pooled type R, C, P and L groups, etc (1023 base-groups) as shown in Table 2. The $MSQ_{(I)}$, $MSQ_{(II)}$ or $MSQ_{(III)}$ values were then calculated for each base-group combination and the same process repeated for each other target group (i.e. C, P, L, O, T, I, U and H) in turn. In this way, a total of 10(1+511+511) = 10230 base-groups were generated according to the project type grouping.

Level 2a - Project specification grouping

For level 2, all the target groups were sub-divided into different base groups according to (1) their preliminary specification type and (2) CFA type as shown in Table 1. In Table 3, for example, the Residential base-group is split into: (i) average standard [Ra], (ii) luxury standard [Rx], (iii) public housing standard [Rp] and (iv) single person public housing standard [Rps]. For an average standard Residential target project Ra, therefore, the base-groups that can be used are: (1) the target base-group (i.e. Ra group only), (2) all combinations of the other three non-target groups (i.e. Rx, Rp and Rps), (3) target group combined with any one to three of the other base groups. The same process is then repeated for all the other target groups (i.e. Rx, Rp and Rps) in turn, after which the whole process is repeated again for the other level one target groups (i.e. C, P, L, O, T, I, U and H).

Level 2b - Project CFA grouping

The level 2b base-groups also involve sub-dividing the target groups (R, C, P, L, O, T, I, U and H) according to their CFA type. In Table 4, for example, the Residential base group is split into: (i) small CFA [RS], (ii) medium CFA [RM], and (iii) large CFA [RL]. Again, for the target project RS, the available base-groups are: (1) the target base-group (RS group only), (2) all combinations of the other two non-target groups (i.e. RM and RL), and (3) target group combined with any one to two of the other base groups. Again, the same process is then repeated for all the other target groups (i.e. RM and RL) in turn, after which the whole process is repeated again for the other level one target groups (i.e. C, P, L, O, T, I, U and H).

IDENTIFYING THE BEST DATA-POOLING ARRANGEMENT

To identify the best data-pooling arrangement, the mean, variance and number of observations in the base-groups were calculated and the formulae (2), (8) and (12) were applied to compute the mean square error values of $MSQ_{(I)}$, $MSQ_{(II)}$ and $MSQ_{(III)}$ for each base group. They were then rank ordered from lowest to highest value and Table 5 provides the results for the first 44 of these. This shows the best base-group to be the RCPLSI pooled project types, with a *MSQ*=6.6937, followed by the RPLSTIU pooled project types, with a *MSQ*=6.6951, in comparison with the relatively poor *MSQ*=6.7122 for the single R target project type base-group.

The results of all the analyses (best pooling arrangement) are summarized in Table 6. This shows all the $MSQ_{(I)}$ results in comparison with the $MSQ_{(III)}$ results (none of the $MSQ_{(II)}$ results improve on the $MSQ_{(I)}$ results). Again, for the Residential results at the project type level, using the base group cost data to conduct the mean value forecast, target project with residential nature TP[R] obtains the $MSQ_{(I)}$ value of 6.7122 while, by using the combined target and non-target group data, the best pooling arrangement of [C,P,L,S,I] provides the best $MSQ_{(III)}$ values of 6.6937 with 0.28% improvement in forecasting. There are a total of 46 out of 511 pooling arrangements which can provide a better $MSQ_{(III)}$ result.

To further the depth of discussion, looking at the Residential results at the project specification level, the MSQ(I) result is 1.2758 by using the TP[Rps], there is 2.5% improvement when using the combined data pool of [Rps,Ra,Rx]. In Table 6, the most outstanding improvement in fit (62.38%) is illustrated in the Hotel target group. The MSQ (I) value (7.665) for TP [HS] under project area level is much improved by using the combine data pool of [TS, TM, TL] with the MSQ (III) value (2.8801).

In general, the Table shows that the $MSQ_{(III)}$ values are often better than $MSQ_{(I)}$ with a suitable pooling arrangement.

CONCLUSIONS

In the very early stages of construction projects, there are situations where only such basic information as project type, project size and preliminary project specification are known concerning a new (target) project and the client/owners' consultant estimators have to resort to either using the price of a very similar project or mean price of a (base) group of projects. This gives rise to an optimisation problem, as enlarging the size of the base-group lessens the variability of the forecast made in this way but, at the same time, also lessens its appropriateness (homogeneity). Following proposals by Skitmore (2001) and Yeung and Skitmore (2005), an empirical method is developed for identifying the base-group that provides the best trade-off of these opposing features. This involves the use of closed form equations to apply the cross validation (leave-one-out) method to generate simulated forecast errors needed to make comparisons between alternative base group compositions. An analysis is described involving 450 actual Hong Kong construction projects in terms of its project type (i.e. Residential, Commercial Centers, Car Parks, Social Community Centres, Schools, Offices, Hotels, Industrial, Universities and Hospitals). This demonstrates the use of the method and what it can achieve. That is, the method identified the set of projects that, by taking the mean of their prices, produce the best simulated estimate of the target project. In the course of this demonstration, it is also shown that using historical data with the same characteristics as the target does not always generate the simulated best forecasts in this situation, with some pooling of data from projects with non-target characteristics usually providing the simulated best forecasts for different targets project.

However, there does not appear to be any noticeable trend in the best pooling arrangements and it seems that these need to be established on a project type by project type basis. The method described here enables this to be done. Furthermore, as the formulae involved need only the number of projects and the mean and variance of their prices to be known for each base-group, the calculations involved are very simple as the only additional work needed is to list the various combinations – a task that can be undertaken easily enough by a spreadsheet or small computer program.

Of course, the analysis provided above is not comprehensive, as it concentrates on only one grouping (project type) and two types of potential project characteristics (GFA and Spec). In practice, other project characteristics may need to be taken into account but it is clear that the proposed method will, however, easily extend to other possible category groupings for further analysis. Another limitation of the method is that it simply uses the mean value of the base-group directly as the forecast while, in conventional practice, values obtained from historical records are often adjusted to some extent by the forecaster's judgement. However, in this sense, the method is no different to that carried out in practice except in the way the initial value is delivered – the forecaster is still free to make any adjustment necessary of the mean value provided.

A final comment is that it may be possible to use the method to provide a more scientific base for the classification of construction cost data. At present, the classification is based primarily on building function (type), with some sub-divisions into large and small (size) and finer measures of function (Spec). The extent to which these existing functional classifications are relevant in terms of classifying project price is not known to have ever been tested empirically (Pegg 2012). For example, the current system places hospitals and schools into two distinct categories due to their different functions – medical and education. However, it is not at all clear that the prices of these two types of building must necessarily be different just because one's function is concerned with medicine and one with education. A more obvious classification would be based on the amount of construction work involved. Market forces, on the other hand, have also been considered to be major determinants of building price (e.g., Skitmore *et al*, 2006), in which case the current classification system may well be appropriate. These two viewpoints are clearly contradictory and the method offers an opportunity to provide a resolution by empirical means.

REFERENCES

Beeston, D.T. (1975) One statistician's view of estimating *Chartered Surveyor: Building and Quantity Surveying Quarterly*, 49-54.

Berkhin, P. (2006). *A survey of clustering data mining techniques: grouping multidimensional data – recent advances in clustering*. ed. Kogan, J., Nicholas, C. and Teboulle, M., 25-71. Netherlands: Springer Berlin Heidelberg.

Flanagan, R. (1980) *Tender price and time prediction for construction work*, PhD Thesis, University of Aston in Birmingham.

Pegg, I. (2012) Personal communication from Chief Statistician. Building Cost Information Service (BCIS).

Skitmore, R.M. (1991) *Early stage construction price forecasting; a review of performance,* The Royal Institution of Chartered Surveyors.

Skitmore, R.M. (2001). Raftery curves for tender price forecasting: empirical probabilities and pooling. *Financial Management of Property and Construction* **6**(3)141-54.

Skitmore, R.M., Runeson, G., Xinling Chang. (2006) Construction price formation: full-cost pricing or neoclassical microeconomic theory? *Construction Management and Economics* **24**(7) 773-84.

Yeung, K.L. (Daisy), Skitmore, R.M. (2005). Data-pooling for early stage price forecasts. *Proceedings*, *21st Annual Conference*, ARCOM, 7-9 Sep, London, SOAS, Vol 1, pp.269-76. Pub. ARCOM, Association of Researchers in Construction Management, c/o School of Construction and Property Management, University of Salford, Maxwell Building, Salford M5 4WT, UK. ISBN 0 902896 93 2.

Table 1. Data clustering table

Level 1: Project type	Level 2a: Specification type	Level 2b: CFA type
Residential (R)	Average standard (Ra)	Residential small CFA (RS)
	Luxury standard (Rx)	Residential medium CFA (RM)
	Public housing standard (Rp)	Residential large CFA (RL)
	Single person public housing standard (Rsp)	
Commercial centre (C)	Average standard (Ca)	Commercial small CFA (CS)
	Luxury standard (Cx)	Commercial medium CFA (CM)
		Commercial large CFA (CL)
Car parking (P)	No specification branch	Car park small CFA (PS)
		Car park medium CFA (PM)
		Car park large CFA (PL)
Social community (L)	No specification branch	Social community small CFA (LS)
		Social community medium CFA (LM)
		Social community large CFA (LL)
School (S)	Primary school (Sp)	School small CFA (SS)
	Secondary school (Ssd)	School medium CFA (SM)
	International school (SI)	School large CFA (SL)
Office (O)	Average standard (Oa)	Office small CFA (OS)
	Luxury standard (Ox)	Office medium CFA (OM)
		Office large CFA (OL)
Hotel (T)	5 Star standard (5ST)	Hotel small CFA (TS)
	3 Star standard (3ST)	Hotel medium CFA (TM)
		Hotel large CFA (TL)
Industrial (I)	No specification branch	Industrial small CFA (IS)
		Industrial medium CFA (IM)
		Industrial large CFA (IL)
University (U)	No specification branch	University small CFA (US)
		University medium CFA (UM)
		University large CFA (UL)
Hospital (H)	No specification branch	Hospital small CFA (HS)
		Hospital medium CFA (HM)
		Hospital large CFA (HL)

Table 2. Level 1: Project Type – Pooling of data groups (for residential target project: Type 1 [R])

Base-group combination	**Type 2 Base-groups:** **All combinations of project type *excluding the target type [R]:***	**Type3 Base-groups:** **All combinations of other project types and the target type [R]:**
any 1 of the project data; $C_1^9 = 9$ (BG)	[C],[P],[L],[S],[O],[T],[I],[U], and [H]	[*R*,C],[*R*,P],[*R*,L],[*R*,S],[*R*,O],[*R*,T],[*R*,I], [*R*,U], [*R*,H]
any 2 of the project data; $C_2^9 = 36$ (BG)	[C,P],[C,L],[C,S],[C,O],[C,T],[C,I],[C,U],[C,H], [P,L],[P,S], … continue to combine any 2 project group data to the last combination of [U,H]	[*R*,C,P],[*R*,C,L],[*R*,C,S],[*R*,C,O],[*R*,C,T],[*R*,C,I],[*R*,C,U] [*R*,C,H], … continue to combine any 2 project group data to the last combination of [R,U,H]
any 3 of the project data; $C_3^9 = 84$ (BG)	[C,P,L],[C,P,S],[C,P,O],[C,P,T],[C,P,I],[C,P,U], [C,L,S] … continue to combine any 3 project group data to the last combination of [I,U,H]	[*R*,C,P,L],[*R*,C,P,S],[*R*,C,P,O],[*R*,C,P,T], [*R*,C,P,I],[R,C,P,U] … continue to combine any 3 project group data to the last combination of [R,I,U,H]
any 4 of the project data; $C_4^9 = 126$ (BG)	[C,P,L,S],[C,P,L,O],[C,P,L,T],[C,P,L,I],[C,P,L, U][C,P,L,H][C,P,S,O] … continue to combine any 4 project group data to the last combination of [T,I,U,H]	[*R*,C,P,L,S],[*R*,C,P,L,O],[*R*,C,P,L,T], [*R*,C,P,L,I],[*R*,C,P,L,U] … continue to combine any 4 project group data to the last combination of [*R*,T,I,U,H]
any 5 of the Project data; $C_5^9 = 126$ (BG)	[C,P,L,S,O],[C,P,L,S,T],[C,P,L,S,I],[C,P,L,S,U], [C,P,L,S,H],… continue to combine any 5 project group data to the last combination of [O,T,I,U,H]	[*R*,C,P,L,S,O],[*R*,C,P,L,S,T],[*R*,C,P,L,S,I], [*R*,C,P,L,S,U],[*R*,C,P,L,S,H],… continue to combine any 5 project group data to the last combination of [R,O,T,I,U,H]
any 6 of the project data; $C_6^9 = 84$ (BG)	[C,P,L,S,O,T],[C,P,L,S,O,I],[C,P,L,S,O,U], [C,P,L,S,O,H][C,P,L,S,T,I],… continue to combine any 6 project group data to the last combination of [S,O,T,I,U,H]	[*R*,C,P,L,S,O,T],[*R*,C,P,L,S,O,I], [*R*,C,P,L,S,O,U], … continue to combine any 6 project group data to the last combination of [*R*,S,O,T,I,U,H]
any 7 of the project data; $C_7^9 = 36$ (BG)	[C,P,L,S,O,T,I],[C,P,L,S,O,T,U],[C,P,L,S,O,T,H], … continue to combine any 6 project group data to the last combination of [L,S,O,T,I,U,H]	[*R*,C,P,L,S,O,T,I],[*R*,C,P,L,S,O,T,U], [*R*,C,P,L,S,O,T,H],… continue to combine any 6 project group data to the last combination of [*R*,L,S,O,T,I,U,H]
any 8 of the project data ; $C_8^9 = 9$ (BG)	[C,P,L,S,O,T,I,U],[C,P,L,S,O,T,I,H], [C,P,L,S,O,T,U,H], … continue to combine any 6 project group data to the last combination of [P,L,S,O,T,I,U,H],	[*R*,C,P,L,S,O,T,I,U],[*R*,C,P,L,S,O,T,I,H], [*R*,C,P,L,S,O,T,U,H]… continue to combine any 6 project group data to the last combination of [*R*,P,L,S,O,T,I,U,H]
9 project data; $C_9^9 = 1(BG)$;	[C,P,L,S,O,T,I,U,H]	[*R*,C,P,L,S,O,T,I,U,H]

Total nos. of base group for all Target Group is: 10x(1[Type 1]+511[Type 2]+511[Type 3])=10,230

Table 3. Level 2a: Project specification – pooling of data groups
(For residential target project Type 1: [Ra],[Rx],[Rp] and [Rsp])

Base-group combination	**Type 2: All combinations of project type *<u>excluding the target type [Ra,[Rx],[Rp] and [Rps]:</u>***	**Type3: All combinations of other project types <u>and the target type [*Ra*],[*Rx*],[*Rp*] and [*Rps*]</u>:**
<u>any 1</u> of the project data; $C_1^3 = 3$ (BG)	[Rx],[Rp],and [Rps]	[*Ra*, Rx],[*Ra*,Rp],and[*Ra*, Rps],
	[Ra],[Rp],and [Rps]	[Rx, Ra],[Rx,Rp],and[*Rx*, Rps],
	[Ra],[Rx],and [Rps]	[*Rp*, Ra],[*Rp*,Rx],and[*Rp*, Rps],
	[Ra],[Rx],and [Rp]	[*Rps*, Ra],[*Rps*,Rx],and[*Rps*, Rp],
<u>any 2</u> of the project data; $C_2^3 = 3$ (BG)	[Rx,Rp],[Rx,Rps],[Rp,Rps]	[*Ra*,Rx,Rp],[*Ra*,Rx,Rps],[*Ra*,Rp,Rps]
	[Ra,Rp],[Ra,Rps],[Rp,Rps]	[*Rx*,Ra,Rp],[*Rx*,Ra,Rps],[*Rx*,Rp,Rps]
	[Ra,Rx],[Ra,Rps],[Ra,Rps]	[*Rp*,Ra,Rx],[*Rp*,Ra,Rps],[*Rp*,Ra,Rps]
	[Ra,Rp],[Ra,Rx],[Rx,Rp]	[*Rps*,Ra,Rp],[*Rps*,Ra,Rx],[*Rps*,Rx,Rp]
<u>any 3</u> of the project data; $C_3^3 = 1$ (BG)	[Rx,Rp,Rps]	[*Ra*,Rx,Rp,Rps]
	[Ra,Rp,Rps]	[*Rx*,Ra,Rp,Rps]
	[Ra,Rx,Rps]	[*Rp*,Ra,Rx,Rps]
	[Ra,Rx,Rp]	[*Rps*,Ra,Rx,Rp]

Total nos. of base group for all Target Group is: 4[Type 1]+4x(7[Type 2]+7[Type 3])=60

Table 4. Level 2b: Project area (CFA) – pooling of data groups

(For residential target project Type 1: [RS],[RM] and [RL])

Base-group combination	**Type 2: All combinations of project type *excluding the target type [RS],[RM]and [RL]:***	**Type3: All combinations of other project types and the target type [*RS*],[*RM*] and [*RL*]:**
any 1 of the project data; $C_1^2 = 2$ (BG)	[RM]and [RL] [RS]and [RL] [RS]and [RM]	[*RS*,RM]and [*RS*,RL] [*RM*,RS]and [*RM*,RL] [*RL*,RS]and [*RL*,RM]
all 2 of the project data; $C_2^2 = 1$ (BG)	[RM,RL] [RS,RL] [RS,RM]	[*RS*,RM,RL] [*RM*,RS,RL] [*RL*,RS,RM]

Total nos. of base group for all Target Groups is: 10X {3[Type 1]+2x(6[Type 2]+3[Type 3])}=210

Table 5: Comparison of mean square values of MSQ(I) and MSQ(III) for the Residential (R) target group

	Target group	Mean square values of target group MSQ(I)	Composition of base-group	Mean square values of base-group MSQ(III)
1	**R**	**6.7122**	**R,C,P,L,S,I**	**6.6937**
2	R	6.7122	R.P.L.S.T.I.U	6.6951
3	R	6.7122	R.L.S.I.U	6.6957
4	R	6.7122	R.P.S.I.U.H	6.6957
5	R	6.7122	R.P.L.S.U.H	6.6959
6	R	6.7122	R.P.L.S.T.I	6.6966
7	R	6.7122	R.P.S.T.I	6.6968
8	R	6.7122	R.P.L.S.T	6.6972
9	R	6.7122	R.P.L.S.I.U.H	6.6979
10	R	6.7122	R.S.I	6.6987
11	R	6.7122	R.L.S	6.6990
12	R	6.7122	R.C.P.L.I	6.7000
13	R	6.7122	R.P.S.U.H	6.7003
14	R	6.7122	R.P.S.H	6.7007
15	R	6.7122	R.P.L.T.I.U	6.7024
16	R	6.7122	R.L.S.I	6.7027
17	R	6.7122	R.S	6.7037
18	R	6.7122	R.S.I.U	6.7037
19	R	6.7122	R.C.P.S.I	6.7037
20	R	6.7122	R.P.S.T	6.7045
21	R	6.7122	R.L.S.U	6.7045
22	R	6.7122	R.C.P.L.S	6.7045
23	R	6.7122	R.P.S.T.I.U	6.7046
24	R	6.7122	RPLSTU	6.7054
25	R	6.7122	RCPLSIU	6.7056
26	R	6.7122	RLIU	6.7062
27	R	6.7122	RPLSH	6.7064
28	R	6.7122	RPSIH	6.7065
29	R	6.7122	RPLSOI	6.7065
30	R	6.7122	RCPI	6.7068
31	R	6.7122	RCPLIU	6.7073
32	R	6.7122	RCPL	6.7075
33	R	6.7122	RPLOI	6.7077
34	R	6.7122	RPTI	6.7078
35	R	6.7122	RPLT	6.7080
36	R	6.7122	RPUH	6.7084
37	R	6.7122	RPTIU	6.7085
38	R	6.7122	RPIUH	6.7087
39	R	6.7122	RIU	6.7088
40	R	6.7122	RPLUH	6.7088
41	R	6.7122	RPLTU	6.7092
42	R	6.7122	RLU	6.7095
43	R	6.7122	RLSIH	6.7098
44	R	6.7122	RPLSTIH	6.7101
45	R	6.7122	RPT	6.7107
46	R	6.7122	RPSU	6.7116

Table 6: Summary of the best pooling arrangement for mean value forecasting

Target group	Mean square error values MSQ(I)	Best combined base group	Mean square error values MSQ(III)	MSQ(I) – MSQ(III)	Improvement in fit (%)	No. of pools provide better forecasting result
R	**6.7122**	**R,C,P,L,S,I**	**6.6937**	**0.0185**	**0.28%**	**46 out of 511**
Rps	**1.2758**	**Rps, Ra, Rx**	**1.2439**	**0.0319**	**2.50%**	**1 out of 28**
RM	5.2036	RM,RL	5.1733	0.0303	0.58%	2 out of 3
C	1.8229	C,O,T	1.7895	0.0334	1.83%	3 out of 511
CS	1.8455	CS,CM,CL	1.5659	0.2796	15.15%	3 out of 3
CM	0.5534	CM,CL	0.5454	0.0080	1.45%	1 out of 3
CL	2.3279	CL,CS,CM	2.2826	0.0453	1.95%	3 out of 3
PM	0.7660	PM,PS,PL	0.6218	0.1442	18.83%	3 out of 3
PL	0.3404	PL,PM	0.3383	0.0021	0.62%	1 out of 3
L	0.5528	L,I	0.5349	0.0179	3.24%	1 out of 511
S	1.1692	S,R,P,O,I,U	1.1567	0.0125	1.07%	29 out of 511
SS	1.4550	SS,SM,SL	1.4224	0.0326	2.24%	3 out of 3
SM	1.0409	SM,SS,SL	1.0180	0.0229	2.20%	3 out of 3
O	2.6419	O,C,U	2.6235	0.0182	0.69%	4 out of 511
OS	2.9137	OS,OM,OL	2.7752	0.1385	4.75%	3 out of 3
OM	1.2409	OM,OS	1.2014	0.0395	3.18%	2 out of 3
OL	3.7536	OL,OS,OM	3.6561	0.0975	2.60%	3 out of 3
T	7.0722	T,H	7.0105	0.0617	0.87%	1 out of 511
TS	**7.6550**	**TS,TM,TL**	**2.8801**	**4.7749**	**62.38%**	**3 out of 3**
I	0.8332	I,L	0.8081	0.0251	3.01%	2 out of 511
IS	1.8465	IS,IM,IL	1.5568	0.2897	15.69%	3 out of 3
IM	0.4579	IM,IS,IL	0.4150	0.0429	9.37%	2 out of 3
IL	0.8695	IL,IM	0.7783	0.0912	10.49%	3 out of 3
U	12.0870	U,C,L,S,O,T,I,H	10.6520	1.4350	11.87%	163 out of 511
US	16.0013	US,UM,UL	15.0190	0.9821	6.14%	3 out of 3
UM	13.8252	UM,US,UL	11.6350	2.1912	15.85%	3 out of 3
H	12.8920	H,T	11.9040	0.988	7.66%	1 out of 511
HS	28.6616	HS,HM,HL	23.2690	5.3925	18.81%	3 out of 3
HM	14.4765	HM,HS,HL	10.4990	3.9771	27.47%	3 out of 3
HL	7.8797	HL,HM	4.8403	3.0394	38.57%	3 out of 3

Figure 1: Flow chart of the analysis process