STATISTICALLY PREPROCESSED DATA BASED PARAMETRIC COST MODEL FOR BUILDING PROJECTS

Sae-Hyun Ji[1], Moonseo Park[2] and Hyun-Soo Lee[3]

[1] Ph.D. Student, Dept. of Architecture, Seoul National University
[2]Professor, Dept. of Architecture, Seoul National University
[3]Professor, Dept. of Architecture, Seoul National University
Correspond to mspark@snu.ac.kr

ABSTRACT: For a construction project to progress smoothly, effective cost estimation is vital, particularly in the conceptual and schematic design stages. In these early phases, despite the fact that initial estimates are highly sensitive to changes in project scope, owners require accurate forecasts which reflect their supplying information. Thus, cost estimators need effective estimation strategies. Practically, parametric cost estimates are the most commonly used method in these initial phases, which utilizes historical cost data (Karshenas 1984, Kirkham 2007). Hence, compilation of historical data regarding appropriate cost variance governing parameters is a prime requirement. However, precedent practice of data mining (data preprocessing) for denoising internal errors or abnormal values is needed before compilation. As an effort to deal with this issue, this research proposed a statistical methodology for data preprocessing and verified that data preprocessing has a positive impact on the enhancement of estimate accuracy and stability. Moreover, Statistically Preprocessed data Based Parametric (SPBP) cost models are developed based on multiple regression equations and verified their effectiveness compared with conventional cost models.

Keywords: Cost, Cost model, Estimate, Data Preprocessing

1. Introduction

Every construction projects have unique characteristics that must be considered during the cost estimating and checking activities. Especially, in the conceptual and schematic design stages, owners require more than accurate cost estimates for information providing. However, information related to the project scope is more likely to change in the early design phases in response to ongoing scope change. Also, as only minimal project scope information is available, during the early stage of estimation, cost estimators need effective estimation strategies.

Generally, cost estimates are based on the estimator's experience, imaginative abilities, and a wide range of assumptions including appraisals of previously conducted projects that are similar in scope (Jarde & Alkass 2007). Practically, parametric cost estimates, developed by adopting regression analysis (Hegazy and Ayed 1998), are most likely to be carried out in the initial phases. One of the most common parametric estimation methods is the unit cost method (e.g., cost per square foot, cost per bed for a hospital), which utilizes either historical building cost data or cost books to obtain an estimate of a building's cost per square foot (Karshenas 1984, Kirkham 2007). Recently, Soutos and Lowe (2005) developed a parametric cost model adopting multiple regression equation based on buildings cost data and identified cost significant variables. In this way, the parametric method is based on historical data collected from similar past projects and scope reflecting parameters (Hendricson 2000). However, the intricate interactions among cost variance impact factors make negative influence on estimate accuracy and employment (Garza and Rouhnan 1995). As demonstrated by previous researches, the compilation of historical data with the cost variance impact parameters is a prime requirement for the preparation of liable and accurate cost estimating. However, despite their arguments on this issue, few approaches were made toward application of data preprocessing on the cost estimate of construction projects. Data preprocessing is a precedent practice of data mining for denoising internal errors or abnormal values. Data mining is the process of extracting useful information from database and looking for useful patterns that can aid decision making (Han & Kamber 2003).

As an effort to deal with this issue, this research develops a statistical methodology that determines cost governing factors and performs data preprocessing. Thereafter, based on preprocessed data, parametric cost estimate equation is developed using multiple regression analysis. To develop this Statistically Preprocessed data Based Parametric (SPBP) cost model, the building cost data of 124 apartment projects in Korea are compiled. The main objectives of the SPBP model are to effectively prepare strategic and conceptual cost estimates, and to provide a check and control functions during the conceptual and schematic design stages. The SPBP cost model is expected to enhance cost estimation field by providing more accurate and stable result than conventional ones.

Previous Researches

Teicholz (1993) argued that a good forecasting method should not require input data that are expensive or difficult to collect. Instead, it should be simple, accurate, unbiased, timely, and stable enough to be easily integrated into the cost system. Ellsworth (1998), on the other hand, argued that the simplest method for determining a reasonable estimate of facility cost is to identify and compare the cost of one project to the cost of similar projects. Furthermore, Ellsworth's research indicates that parametric estimating methods—predominantly used by owners at the conceptual stage of construction—can be an effective cost estimating strategy in which the characteristics of similar projects are used as parameters.

Traditionally, parametric cost estimating models are developed by applying regression analysis to historical project information. Yet, it is difficult to develop these models due to the inherent limitations of regression analysis (Hegazy and Ayed 1998). One major disadvantage of regression based techniques is that they require a defined mathematical form for the cost function that best fits to available historical data (Creese and Li 1995). Another disadvantage is their unsuitability to account for a large number of variables involved in construction projects and knotty intricate interactions among these variables. These limitations contribute to the low accuracy of traditional models, and to their limited use in construction (Garza and Rouhnan 1995).

However, previous research rarely examined data preprocessing as a viable component of cost estimating. In the literature, the cost estimating formula has been deduced using regression analysis based on survey results that rank parameters on nominal and ordinal scales (Chan & Park 2005), despite the fact that these parameters consist of real technical and qualitative data. Cheng et al. (2008) have developed a web-based conceptual cost estimate system by introducing techniques such as genetic algorithms, fuzzy logic, and neural networks. Nevertheless, as their system uses many impact factors and requires a complicated process, it is likely that the system will be regarded as a "black box." Soutos and Lowe (2005), on the other hand, developed a parametric cost model using regression analysis based on a database of 360 buildings. However, their research did not deal with data preprocessing and did not address the presence of multicollinearity.

Furthermore, Trost and Oberlender (2003), using multivariate regression analysis based on the results of factor analysis, identified five factors that exhibit a significant impact on estimate accuracy. Their research indicates that the drivers of estimate accuracy must be adequately identified and quantified. And alos, Hegazy and Ayed (1998) developed a parametric cost estimating model to address the lack of project scope information. Their model adopts the neural networks approach and consists of one or more functions, or cost estimating relationships, between cost (as a dependent variable) and the cost determining factors (as independent variables). This approach is appropriate because it can address difficult tasks that require intuitive judgment, and it can detect data patterns that elude conventional analytical techniques.

By utilizing data on the physical characteristics of parameters, parametric methods allow the conversion of technical values to quantitative economic data. Because, the data may have "noise" that can have a negative impact on the reliability or accuracy of the estimates, data cleansing—which is one of the preprocessing technique and eliminates internal errors or abnormal values in the data—needs to be performed, before using them for estimating. Also, normalization—which in statistical terms refers to the division of multiple data sets by a common variable in order to negate that variable's effect on the data—should precede data cleansing, before inputting them the process of estimating.

Component Methodology

Principle Component Analysis (PCA)

Principle Component Analysis (PCA) is the simplest eigenvector-based multivariate analyses. Its operation is often thought of as revealing the internal structure of data in a way that best explains variance in that data (Kwan 2008). PCA is a technique for forming new variables which are linear composites of the original variables. The new variables are called principle components and the values of the new variables are called principle components scores. The first principle component accounts for the maximum variance in the data. PCA, which supplies a lower-dimensional picture, is commonly referred as a data reduction technique without losing characteristics (Sharma 1996). If there might be high correlation among variables, or one to two high variance variables in the data set, most of the information of original data could be represented by the first components. Thus, the contribution of variables can be prioritized with principle component scores (Kwan 2008).

Multiple regression analysis

Regression analysis is used when the dependent variable and the multiple independent variables of the model are measured using a metric scale resulting in metric data. Regression analysis is generally used for prediction, hypothesis testing, and modeling of causal relationships. As multiple regression analysis contains many independent variables, it is more comprehensive, provides more accurate estimation, and also reduces estimate error. However, using multiple regression analysis usually involves the assumption that multicollinearity and interaction effect do not exist. More commonly, the issue of multicollinearity arises when multiple regression analysis is used which have a high degree of correlation (either positive or negative) between two or more independent variables. In the presence of multicollinearity in the data, stepwise approach may be appropriate for excluding the effect.

Parametric estimation

Parametric estimation has been considered as an effective method by reducing time spent bidding on a job for the construction industry (Bajaj et al. 2002). As parametric estimation takes little time at minimal expense, it is an extremely appealing method for determining approximate project cost. As its name implies, the parametric method is based on certain parameters that are commonly used in the construction industry and that reflect physical project characteristics including size, building type, roof type, exterior closure type, and number of floors. Usually, due to the fact that parametric estimates are prepared before a facility is designed, they are based on historical data collected from similar past projects (Hendricson 2000). The parametric cost model an useful tool not only for preparing early conceptual estimates when there are limited technical data or engineering deliverables (Dysert 2001), it can also help avoid errors and omissions that are common in traditional cost estimating procedures, particularly during the planning and early design phases (Meyer & Burns 1999).

Traditionally, cost estimating relationships are developed by applying regression analysis to historical project information. This study will focus on comparing two commonly used single regression analysis methods: (1) the square foot estimating method, which is based on cost estimating equations that use gross floor area as a parameter, and (2) the unit estimating, the number of households based formulas.

SPBP Cost Modeling

To address the aforementioned problems, the Statistically Preprocessed Data Based Parametric (SPBP) cost model is developed in this research. First, the scope of the cost model is defined. Then, historical data are collected, and within the previously defined cost model scope, the collected data are normalized for time by using the historical cost index. Then, statistical preprocessing of data cleansing for denoising is conducted based on interval estimation sampling. Consequently, the statistically preprocessed data set is abstracted under the unit price, which is adjusted by the first cost variance governing (the dominant) parameter. Using the cleansed data, cost estimate relationships are then derived through a stepwise multiple regression analysis for excluding the effects of multicollinearity. Finally, the SPBP model is validated through a comparative study with other cases. The entire SPBP cost modeling process is diagramed in Fig. 1.

Model Scoping

Cost estimation is crucial to construction contract tendering, providing a basis for establishing the likely cost of resource of the tender price for construction work (Akintoye 2000). The recent researches focus on estimates generated during the initial stages of a project. As such, the primary function of initial estimating is to produce a forecast of the probable cost of a future projects, before the building has been designed in detail and contract particulars are prepared (Seely 1996). Smith (1995) observes that the initial estimate is implemented for review has a particularly significant role because it is the basis for the release of funds for further studies of estimates, and because it becomes the marker against which all subsequent estimates are compared. Moreover, performance and overall project success are often measured by how well the actual cost compares to the early cost estimates (Oberlender & Trost 2001). In this context, the use of the SPBP cost model is to prepare strategic and conceptual estimates for budgeting that can provide the function of cost check and control for the conceptual and design stages.

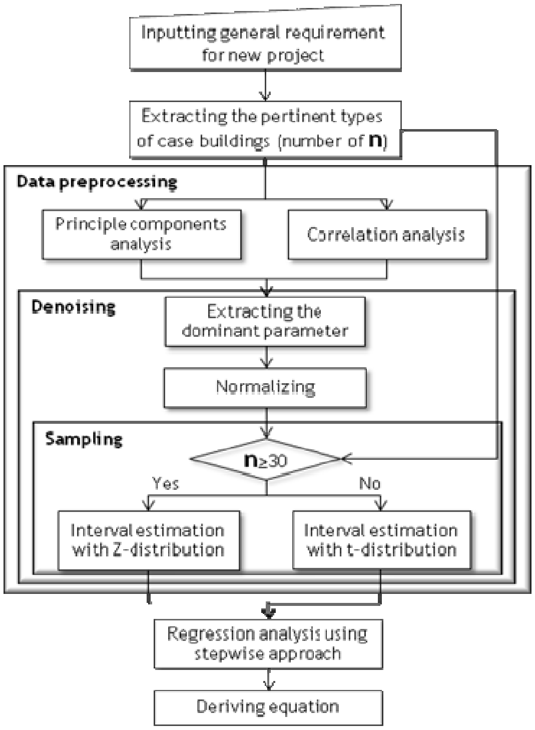
Inputting general requirement for new project
Extracting the pertinent types of case buildings (number of n)
Data preprocessing
Principle components analysis
Correlation analysis
Denoising
Extracting the dominant parameter
Normalizing
Sampling
n≥30
Yes
No
Interval estimation with Z-distribution
Interval estimation with t-distribution
Regression analysis using stepwise approach
Deriving equation

Fig. 1. The SPBP Cost Modeling.

Data Analysis

Currently, in Korea, different types of apartment households are typically developed and produced by unit gross area because these developments have been seriously affected by housing supply legislation. Other apartment building project data related to cost can also be analyzed and/or categorized by the same criteria (e.g, by unit gross area). Consequently, the results of a data analysis are expected to yield similar patterns. For the purpose of properly capturing all the parameters in the cost database (built for the purposes of this study), and subsequently in the SPBP model, the building cost data (priced bills of quantities) of 124 apartment buildings from 11 housing complex projects in Korea are collected. The data are supplied by Seoul Housing Corporation which is a public enterprise established by Korean government. 102 of 2005 year cases are used for model building and 22 of 2007 cases were used for model validation. Then, the collected data should be normalized in respect to escalation, regional location, and system specification. However, in this research, only the data regarding escalation was normalized. This normalization was performed using the historical cost index of 1.025 for converting data of 2005 year to 2007 year—published the Korean government—of each building type. Because of Korea's relatively small territory, there is little point in normalizing the data for regional location, and system specification.

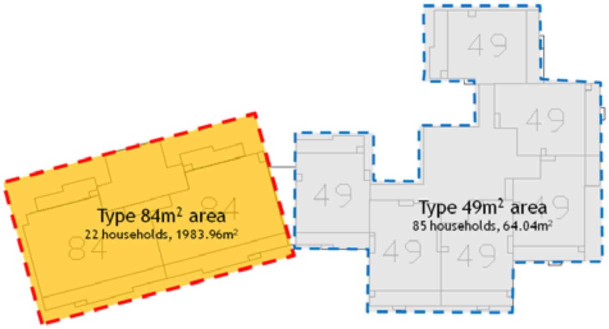

Fig. 2. Example of sorting by type

Then, a cost database is built. It is divided into three categories: cost information, impact factors, and other information. For practicality, some of the impact factors are chosen as parameters for regression equations. As seen in Table 2, all of this information is accumulated and analyzed in the cost database for every single area type: type 84(m2), 59(m2), and 49(m2). Through a comprehensive analysis of the plan of apartment buildings, it is determined that all the buildings are either the singular type of unit gross area or a combination of plural types. Type is a set of same unit area households. Thus, all historical cost data classified to make the singular type of unit gross area are classified independently. On closer examination, the cost data is structured with trades bundled by the sum of quantities of each item (bills of quantities). Accordingly, a cost data analysis is conducted by adapting the gross area ratio into takeoff quantities of each item.

Table 2. Configuration of Cost Database

Category	Factors / Information
Parameter	(X1) Number of households (X2) Gross floor area (X3) Number of unit floor households (X4) Number of elevators (X5) Number of floors (X6) Number of piloti with household scale (X7) Number of households of unit floor per elevator
Cost information	Total cost and unit cost per each impact factor
Other information	Building number, Structure type, Floor height, Number of pit floors, Pit height, Roof type, Piloti type, Piloti height, Top floor type, Shape of floor plan, and Year of design.

Data Preprocessing

Data mining is the process of discovering meaningful patterns and relationship that lie hidden within large quantity of information (Han & Kamber 2003). Data preprocessing is a commonly used preliminary data mining practice, includes any type of processing procedure that acts on and prepares raw data for another processing procedure. Data preprocessing does this by converting the data into a structured format that can be more easily and effectively processed for depending on the user's specific purposes. There are different methods used for data preprocessing. Sampling selects a representative subset from a large population; Transformation, which manipulates raw data to produce a single input. Denoising removes "noise" from data. Normalization organizes data for more efficient access. And Feature extraction pulls out specified data that are significant in a particular context (Han & Kamber 2003). Construction project's cost data normally have "noise," such as internal errors or abnormal values that can have a negative effect on the reliability confidence and accuracy of the estimating result. To address this problem, a data preprocessing process is developed. In this process, the dominant cost variance governing parameter is determined based on PCA, and then normalization is conducted with the dominant parameter, and denoising is performed with sampling applied by interval estimating. In observing the impact factors of the cost database, it is intuitive that some factors will be highly correlated to cost, while also demonstrating strong correlations amongst themselves. The correlation coefficient indicates the strength and direction of a linear relationship between two random variables. Generally, a correlation coefficient of under -0.5 or over 0.5 indicates that the two variables have a strong correlation. In fact, a high bivariate correlation is identified using Statistical Package for the Social Science

Table 3. Correlation Analysis (Pearson Correlation Coefficient)

	Type 49	Type 59	Type 84

	X1	X2	X3	X4	X5	X6	X7	X1	X2	X3	X4	X5	X6	X7	X1	X2	X3	X4	X5	X6	X7
X1	1.00							1.00							1.00						
X2	0.94	1.00						0.98	1.00						0.98	1.00					
X3	0.76	0.75	1.00					0.70	0.67	1.00					0.81	0.76	1.00				
X4	0.00	0.17	0.56	1.00				0.06	0.02	0.51	1.00				0.04	-0.04	0.32	1.00			
X5	0.32	0.30	-0.36	-0.72	1.00			0.70	0.70	0.03	-0.34	1.00			0.53	0.56	-0.02	-0.38	1.00		
X6	0.71	0.72	0.83	0.32	-0.17	1.00		0.32	0.37	0.10	-0.42	0.39	1.00		0.33	0.30	0.38	-0.23	0.21	1.00	
X7	0.83	0.65	0.52	-0.42	0.35	0.57	1.00	0.63	0.65	0.46	-0.53	0.38	0.53	1.00	0.64	0.67	0.55	-0.61	0.32	0.52	1.00
cost	0.92	0.98	0.73	0.15	0.30	0.75	0.64	0.96	0.99	0.73	0.09	0.63	0.39	0.62	0.93	0.96	0.76	-0.07	0.49	0.31	0.70

(SPSS) correlation function between building cost and each of the following factors: gross floor area, number of households, and number of floors. A high correlation is also identified amongst the factors themselves which means there are probability of multicollinearity (Table 3). Using the factor analysis function of SPSS with the correlation matrix extraction approach in the SPBP, two components are extracted from eight variables within type 49 and type 59, and three components are extracted for type 84. The first components of PCA can represent most of the information of the original data. The percentage of variance, explained by the first component, is over 57% for all types. More precisely, X1 and X2 have the highest coefficient (over 0.95) of the component matrix (Table 4). Thus, X2 (gross floor area) is selected for the dominant parameter for all three building types because its correlation coefficient to cost is higher than X1.

Based on confidence intervals, interval estimation is described with an upper and lower limit and used to indicate the accuracy of an estimate. With the assumption that unit cost data, normalized by gross floor area ($/m2), are approximated by the normal distribution, normalization and statistical sampling using interval estimation is conducted. As a result, it is found that there are 25 cases out of 49 with a 95% Confidence Level (CL) and 30 out of 49 with a 99% CL in type 84; 13 out of 25 with a 95% CL and 9 out of 25 with a 99% CL in type 59; and 7 out of 16 with a 95% CL and 9 out of 16 with a 99% CL in type 49. These cases are selected to be used in deriving the final cost estimating equations (Table 5). A common rule of thumb from the Central Limit Theorem (CLT) is that a normal distribution can be used for approximation when n≥30. Accordingly, there are over 30 cases for type 84 the Z – distribution was used for them, and in case of type 49 and 59 which are less than 30 cases respectively, the t - distribution was applied to these types.

Table 4. Principle Components Scores

	Components (Type 49)		Components (Type 59)		Components (Type 84)		
	1	2	1	2	1	2	3
X1	0.968	-0.190	0.967	0.166	0.956	0.201	-0.164
X2	0.954	-0.072	0.978	0.120	0.962	0.132	-0.202
X3	0.855	0.478	0.669	0.637	0.781	0.534	0.291
X4	0.161	0.936	-0.096	0.954	-0.176	0.946	-0.080
X5	0.152	-0.915	0.698	-0.361	0.542	-0.457	-0.594
X6	0.850	0.284	0.509	-0.530	0.493	-0.244	0.627
X7	0.768	-0.441	0.760	-0.355	0.807	-0.389	0.314
Cost	0.949	-0.074	0.967	0.286	0.952	0.118	-0.139
Eigen values	4.838	2.263	4.605	1.930	4.575	1.672	1.022
% of variance	60.475	28.284	57.567	24.230	57.187	20.902	12.776
Cumulative %	60.475	88.759	57.567	81.687	57.187	78.089	90.865

Table 5. Results of Interval Estimation for Each Confidence Level

Type	49	59	84

Confidence level	100%	99%	95%	100%	99%	95%	100%	99%	95%
Mean	476.71	476.81	473.71	565.22	552.09	556.86	531.65	527.75	531.78
Standard deviation	14.69	4.63	3.10	57.74	16.32	14.07	66.89	11.48	8.81
Lower limit	457.20	464.63	467.49	450.31	529.52	537.60	470.94	507.04	512.92
Upper limit	508.38	488.78	485.93	708.28	600.91	592.83	928.17	556.26	550.38
Number of cases	16	9	7	25	17	13	49	30	25

Table 6. Parametric Equations Using Stepwise Multiple Regression

Type	Confidence level	Equations	Standard errors	t-statistics	Significance level	Adjusted R^2
49	100%	= – 52188 + 487.680 * X2	74108	16.484(X2)	0.000(X2)	0.948
	99%	= – 106914 + 495.272 * X2	21633	21.584(X2)	0.000(X2)	0.985
	95%	= – 91524 + 493.809 * X2	13384	34.735(X2)	0.000(X2)	0.995

59	100%	= 311396 + 547.368 * X2 – 26698.939 * X5	113245	26.359(X2)
				–2.878(X5)
					0.000(X2)
					0.009(X5)
						0.980

	99%	= – 9972 + 484.533 * X2 + 68708.898 * X3	45159	25.492(X2)
				3.016(X3)
					0.000(X2)
					0.010(X3)
						0.996

	95%	= – 10372 + 492.361 * X2 + 63810.302 * X3	36666	26.483(X2)
				2.687(X3)
					0.000(X2)
					0.023(X3)
						0.998

84	100%	= 250069 + 470.656 * X2	182810	23.991(X2)	0.000(X2)	0.922
	99%	= – 6665 + 529.425 * X2	55567	60.087(X2)	0.000(X2)	0.992
	95%	= – 16091 + 531.382 * X2	42507	71.992(X2)	0.000(X2)	0.995

Deriving Cost Estimating Relationships

To determine cost estimating relationships, multiple regression analysis —a statistical process for analyzing the relationship between quantitative variables—is utilized. The multiple regression method is traditionally adopted for the purpose of developing parametric cost estimating equations, its main objective is to build a mathematical formula that will assist in accurately predicting the impact on a variable when a related variable changes. However, to prevent the presence of multicollinearity, the stepwise approach is applied using regression function of SPSS. Stepwise approach is a combination of forward and backward procedures. Forward begins with no independent variables in the equation. The high correlated variable is entered into the equation first. The rest are entered into the equation depending on their contribution. On the contrary backward begins with all variables in the equation and sequentially removes them (Park et al. 2004). Table 6 reports the regression equations using preprocessed data. These equations have different confidence levels and different estimated effects of the individual variable on building cost.

Table 7. Comparison of Accuracy Ratio (SPBP vs. Traditional method)

Type				parameter				SPBP equation with CL of			$/GFA	$/NH
	X1	X2	X3	X4	X5	X6	X7	100%	99%	95%		
59	19	1597.1	2	1	10	0	2	126.21%	123.83%	124.15%	124.02%	123.93%
	24	1992.5	2	1	12	0	2	117.72%	118.94%	119.53%	122.57%	124.01%
	26	2188.4	2	1	14	2	2	118.19%	123.64%	124.36%	128.75%	128.48%
	28	2321.5	2	1	14	0	2	124.25%	128.77%	129.59%	134.93%	136.69%
	30	2492.3	2	1	15	0	2	116.21%	121.68%	122.52%	128.39%	129.81%
	38	3185.2	4	2	10	2	2	103.56%	104.74%	105.03%	104.29%	104.50%
	56	4420.3	4	1	15	2	4	97.61%	100.80%	101.42%	104.65%	111.36%
	58	4565.1	4	1	15	0	4	99.66%	102.43%	103.08%	106.71%	113.88%
	60	4687.7	4	1	15	0	4	102.55%	105.01%	105.71%	109.71%	117.95%
	60	4687.7	4	1	15	0	4	110.83%	113.48%	114.23%	118.56%	127.46%
	60	4703.3	4	1	15	0	4	100.34%	102.70%	103.38%	107.32%	115.00%
				Mean				110.65%	113.28%	113.91%	117.26%	121.19%
				Variance				1.08%	1.09%	1.10%	1.24%	0.90%
84	16	1809.1	2	1	9	2	2	122.65%	105.91%	105.25%	107.09%	102.96%
	18	2021.1	2	1	9	0	2	125.91%	111.46%	110.88%	112.62%	109.04%
	18	2024.8	2	1	9	0	2	128.22%	113.55%	112.96%	114.73%	110.88%
	20	2189.9	2	1	10	0	2	130.62%	117.57%	117.04%	118.74%	117.89%
	20	2244.7	2	1	10	0	2	126.22%	114.17%	113.68%	115.29%	111.67%
	22	2410.4	2	1	11	0	2	121.68%	111.57%	111.15%	112.62%	111.75%
	24	2627.2	2	1	12	0	2	122.70%	114.26%	113.90%	115.29%	114.50%
	24	2655.4	2	1	13	2	2	109.23%	101.90%	101.59%	102.81%	101.02%
	26	2879.9	2	1	14	2	2	126.99%	120.08%	119.77%	121.11%	118.86%
	26	2955.7	2	1	14	0	2	126.13%	119.75%	119.47%	120.76%	115.49%
	28	3091.4	2	1	15	2	2	106.66%	101.97%	101.76%	102.81%	101.23%
				Mean				122.46%	112.02%	111.59%	113.08%	110.48%
				Variance				0.59%	0.41%	0.41%	0.42%	0.40%