International Journal of Business, Human and Social Sciences
ISSN: 2517-9411
Vol:2, No:6, 2008

Application of Data Mining Tools to Predicate Completion Time of a Project

Seyed Hossein Iranmanesh, and Zahra Mokhtari

***Abstract*—**Estimation time and cost of work completion in a project and follow up them during execution are contributors to success or fail of a project, and is very important for project management team. Delivering on time and within budgeted cost needs to well managing and controlling the projects. To dealing with complex task of controlling and modifying the baseline project schedule during execution, earned value management systems have been set up and widely used to measure and communicate the real physical progress of a project. But it often fails to predict the total duration of the project. In this paper data mining techniques is used predicting the total project duration in term of Time Estimate At Completion-EAC (t). For this purpose, we have used a project with 90 activities, it has updated day by day. Then, it is used regular indexes in literature and applied *Earned Duration Method* to calculate time estimate at completion and set these as input data for prediction and specifying the major parameters among them using Clem software. By using data mining, the effective parameters on EAC and the relationship between them could be extracted and it is very useful to manage a project with minimum delay risks. As we state, this could be a simple, safe and applicable method in prediction the completion time of a project during execution.

***Keywords*—**Data Mining Techniques, Earned Duration Method, Earned Value, Estimate At Completion.

I. Introduction

A. Earned Value Management System

THIS term of estimate at completion-EAC, is usually used in project management terminologies, is referred to the concept of prediction the time and cost of work completion in a project. A systematic approach for EAC and controlling the baseline project scheduling during execution is earned value management system-EVMS. Earned value management is a methodology to measure and communicate the real physical progress of a project. It also takes into account the work complete, the time taken and the cost incurred to complete the project. It helps to project management team to evaluate and control project risk by measuring project progress in monetary terms. The Project Management Institute defines earned value management (EVM) as “a management methodology for integrating scope, schedule, and resources, and for objectively measuring project performance and progress. Performance is measured by determining the budgeted cost of the work performed (i.e. earned value) and comparing it to the actual cost of the work performed (i.e. actual cost). Progress is measured by comparing the earned value to the planned value.” [8]. The Association for Project Management defines EVM as “a project control process based on a structured approach to planning, cost collection and performance measurement. It facilitates the integration of project scope, time and cost objectives and the establishment of a baseline plan for performance measurement.” [8].

Importance of EVMS in measuring project progress and calculating Earned Value (EV) of project and forecasting EAC could be evident, since correct and on time EAC is very important to plan preventive actions during the project life cycle. This research apply data mining tools to forecast EAC and for this use Clem 8.1 software to reach reasonable results. The basic data for this study was produced by using Progress Project Simulator software. The specification of this simulator is given in the following section.

B. Literature Review

The basic concept in EVM and use it in practice have been comprehensively described in many sources e.g. [3] and [6]. Although EVM has been set up to follow-up both time and cost, many of the research have been focused on the cost aspect; e.g. [7] have tried to introduce two new indexes for forecasting and implement them in some project. [4] have introduced a different equation which is frequently used in ecology and classical project S-curve for forecasting. [1] represented a new formalism and a corresponding new notation for earned value analysis. There are a few papers with a pure focus on the EV in the literature and the literature review shows that growth of EV’s scientific papers has been very slow. In this paper we focus on time estimate at completion-EAC (t). For this purpose the schedule performance measure need to be translated from monetary units to time units. There are three methods in the literature that have been proposed to measure schedule performance: “the planned value method [3], the earned duration method (Jacob & Kanen (2004)) and the earned schedule method that has been recently introduced by (Lipe (2003)). The earned duration method translate the well known SV and SPI indicators from monetary units to time units, and the earned scheduled method calculates two alternative schedule performance measures (referred to as SV(t) and SPI(t)) that are directly expressed in time units.

Seyed Hossein Iranmanesh, Assistant Professor, is with "University of Tehran" & "Institute for Trade Studies & Research", Tehran, Iran (corresponding author, phone: +9821-88021067, fax: +9821- 88013102, e-mail: hiranmanesh@ut.ac.ir).

Zahra Mokhtari is with Socio-Economic Systems Engineering, University of Tehran, Tehran, Iran (e-mail: zmokhtari@ut.ac.ir).

C. *Modeling Input Parameters*

EAC in time and cost could be calculated based on three variables: Budget Cost of Work Scheduled (BCWS) defined as baseline cost scheduled for a project, Budget cost of work performed (BCWP) is scheduled cost for work performed and Actual Cost of Work Performed (ACWP).

Earned value management requires the three following key parameters to measure project performance:

PV Planned Value (BCWS)
AC Actual Cost (ACWP)
EV Earned Value (BCWP)

Project performance in literature, both in terms of time and cost, is determined by comparing the three key parameters PV, AC and EV, resulting in four well-known performance measure:

SV Schedule Variance (SV=EV-PV) (1)
SPI Schedule Performance Measure (SPI=EV/PV) (2)
CV Cost Variance (CV=EV-AC) (3)
CPI Cost Performance Index (CPI=EV/AC) (4)

If CPI were less than 1, the project would completed with a cost higher than scheduled, and if it were equal to 1 the project would be completed on planned cost. In similar way, SPI less than 1 means that the project would be completed more than planned time, and if it were equal to 1, it means that the project would be completed on planned time. But any result couldn't be give if they were more than 1. The cost performance indexes and their application to forecast the final cost of project have been discussed extensively in literature. In this paper we use above indicators and select suitable method to transforming this indexes from cost units to time units. As mentioned before, there are three methods for EVM measuring. The planned value method of [3] relies on the well-known earned value metrics to forecast a project's duration using the following metrics:

PVR Planned Value Rate (=BAC/PD) that BAC is defined as Budget At Completion and PD is Planned Duration (5)
TV Time Variance (=SV/PVR) (6)

Second method is introduced by Jacob and Kane (2004), the term earned duration ED defined as the product of the actual duration and the SPI:

ED Earned Duration (=AD*SPI) (7)

For the purpose of this paper we use EAC (t) as Time Estimate At Completion as follow:

EAC (t) =AD+ (PD-ED)/(CPI*SPI) (8)

Because our purpose is not choosing best formula for estimating end duration of project, we select this formula for time estimating to only represent application of data mining in this field.

Third method is defined by Lipke (2003), *earned scheduled method* relies on similar principles of the earned value method as follow:

Find t such that EV≥PV_t and EV≤PV_{t+1}

ES=t+ (EV- $PV_t)/(PV_{t+1} - PV_t)$ (9)

Such that

ES Earned Schedule

EV Earned value at the actual time

PV_t Planned Value at time instance t

II. Methodology

Data mining is a knowledge discovery technique that is widely used in real word problems. According to the Gartner Group, "Data mining is the process of discovering meaningful new correlations, patterns and trends by sifting through large amounts of data stored in repositories, using pattern recognition technologies as well as statistical and mathematical techniques." The tasks of classification, estimation, prediction, affinity grouping or association rule and clustering could be performed with data mining tools. These tools composed of: Decision Tree, Market Basket Analysis and Association Rules, Clustering, Customer Life Cycle, Artificial Neural Network and Genetic Algorithm [9]. As it was mentioned above, the main goal of this paper is to present an approach for prediction of real duration of a project. To access this purpose, we use some tools of data mining such as: Decision Tree, Neural Network and Association Rule. Decision Tree is an attractive method for clustering and it consists of three steps: preparing data, model building and generating decision rules. Association rule is an approach that widely used in Market Basket Analysis. This method represents specific number of rules that their support and confidence exceed from minimum support and confidence. Artificial Neural network are established with inspiration of neural network in human body. Since neural networks produce continuous output, they may quite naturally be used for estimation and prediction. In this study we use Clem software for representing outputs and sensitivity analysis. The input parameters are: ACWP, BCWP, BCWS, AD, ED, PD, EAC (t).

III. Results

In this study, it is used a sample project from Kulish – Hartmann data set (j303_10) from http://129.187.106.231/psplib/main.html. This data was created for solving Resource Constrained Project Schedule Problem (RCPSP). For more information about RCPS problem and data set, the reader could refer to reference [5]. It is calculated CPI, SPI, ACWP, BCWP, BCWS, AD, ED, PD, EAC (t) for J303_10 project for 76 time periods and it is shown in appendix 1. These data is generated randomly by some basic assumptions which are given in reference [5] for each time period.

PD, ED, CPI and SPI indexes are main indicators for input parameters and Time Estimate At Completion-EAC (t) is assumed as output parameter in Clem software. This information read by Clem with diagram shown in Fig. 1. In this figure INPUT field contains of input variables, and each path shows one of used methods in this study. The results of using these methods could be seen in Figs. 2,3,4,5.

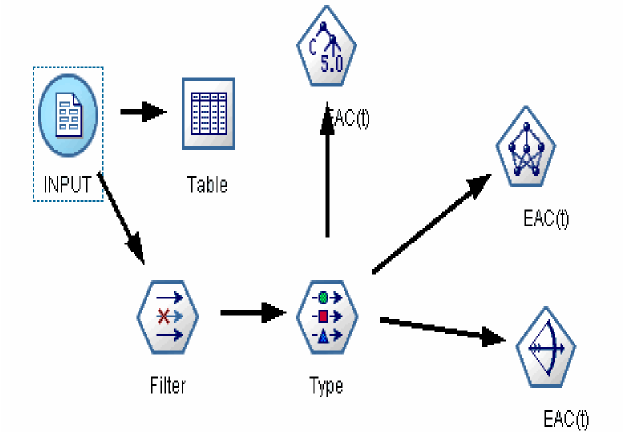
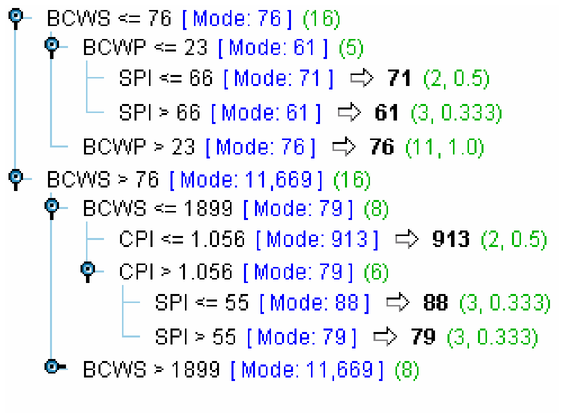
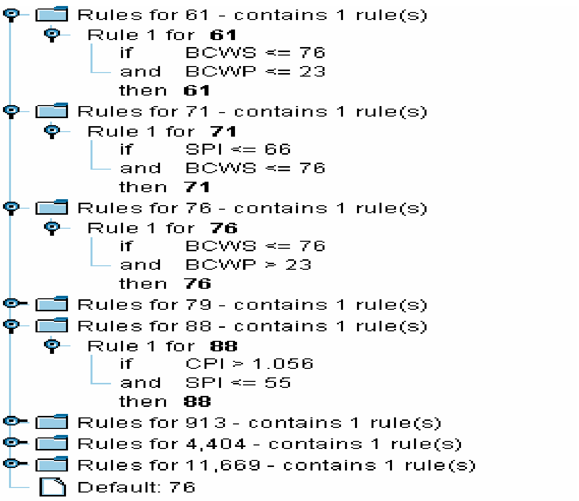


Fig. 1 Diagram of our problem



```
BCWS <= 76 [ Mode: 76 ] (16)
   BCWP <= 23 [ Mode: 61 ] (5)
      SPI <= 66 [ Mode: 71 ]  ⇨  71  (2, 0.5)
      SPI > 66 [ Mode: 61 ]  ⇨  61  (3, 0.333)
   BCWP > 23 [ Mode: 76 ]  ⇨  76  (11, 1.0)
BCWS > 76 [ Mode: 11,669 ] (16)
   BCWS <= 1899 [ Mode: 79 ] (8)
      CPI <= 1.056 [ Mode: 913 ]  ⇨  913  (2, 0.5)
      CPI > 1.056 [ Mode: 79 ] (6)
         SPI <= 55 [ Mode: 88 ]  ⇨  88  (3, 0.333)
         SPI > 55 [ Mode: 79 ]  ⇨  79  (3, 0.333)
   BCWS > 1899 [ Mode: 11,669 ] (8)
```

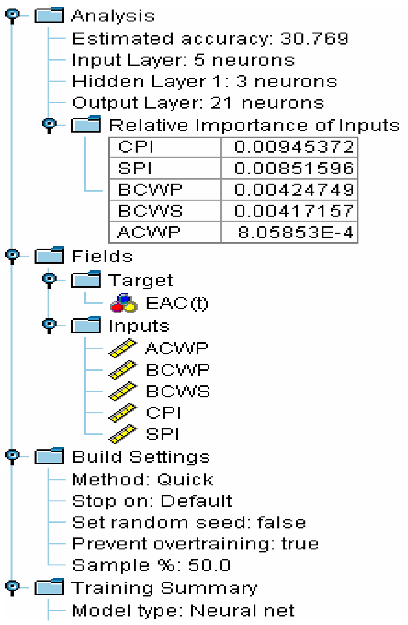
Fig. 2 Decision Tree for input parameters



```
Rules for 61 - contains 1 rule(s)
   Rule 1 for  61
      if      BCWS <= 76
      and     BCWP <= 23
      then    61
Rules for 71 - contains 1 rule(s)
   Rule 1 for  71
      if      SPI <= 66
      and     BCWS <= 76
      then    71
Rules for 76 - contains 1 rule(s)
   Rule 1 for  76
      if      BCWS <= 76
      and     BCWP > 23
      then    76
Rules for 79 - contains 1 rule(s)
Rules for 88 - contains 1 rule(s)
   Rule 1 for  88
      if      CPI > 1.056
      and     SPI <= 55
      then    88
Rules for 913 - contains 1 rule(s)
Rules for 4,404 - contains 1 rule(s)
Rules for 11,669 - contains 1 rule(s)
Default: 76
```

Fig. 3 Rules from decision trees method

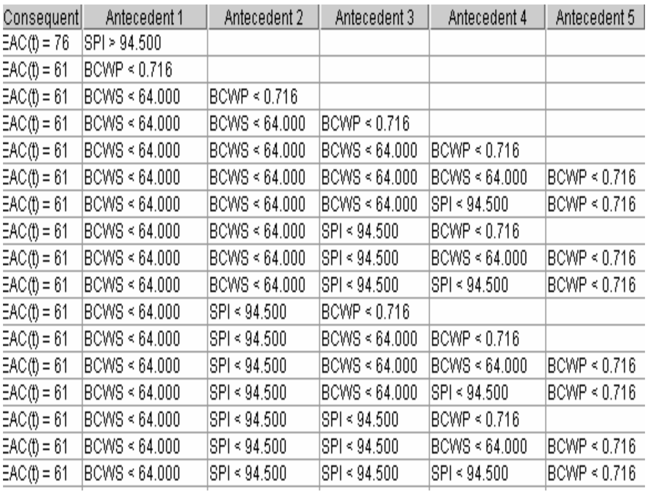
In Fig. 2, BCWS represents the root node split for this problem. For more explain, the first branch indicate that if BCWS<76 and BCWP<23 and SPI<66, the estimate of project will be about 71 time units (days). Also the number in parenthesis inform there are 2 time periods of 71 reporting periods contain these properties with 20% confidence that confirm that EAC will be as 71 days. Each other branches could be interpreted like this. In Fig. 3, rules from decision tree's method execution are represented. For example, rule 1 show that, if BCWS is equal or less than 76 and BCWP is less or equal than 23, then the project duration will be 61 days.



```
Analysis
   Estimated accuracy: 30.769
   Input Layer: 5 neurons
   Hidden Layer 1: 3 neurons
   Output Layer: 21 neurons
   Relative Importance of Inputs
      CPI     0.00945372
      SPI     0.00851596
      BCWP    0.00424749
      BCWS    0.00417157
      ACWP    8.05853E-4
Fields
   Target
      EAC(t)
   Inputs
      ACWP
      BCWP
      BCWS
      CPI
      SPI
Build Settings
   Method: Quick
   Stop on: Default
   Set random seed: false
   Prevent overtraining: true
   Sample %: 50.0
Training Summary
   Model type: Neural net
```

Fig. 4 Neural Network

In Fig. 4 we can see that, CPI has the largest weight among all indexes to predict completion time in a system with neural network predictor. Target and input parameters and the priority of inputs in prediction are shown in this graph ("Relative Importance of Inputs" section).



Consequent	Antecedent 1	Antecedent 2	Antecedent 3	Antecedent 4	Antecedent 5
EAC(t) = 76	SPI > 94.500				
EAC(t) = 61	BCWP < 0.716				
EAC(t) = 61	BCWS < 64.000	BCWP < 0.716			
EAC(t) = 61	BCWS < 64.000	BCWS < 64.000	BCWP < 0.716		
EAC(t) = 61	BCWS < 64.000	BCWS < 64.000	BCWS < 64.000	BCWP < 0.716	
EAC(t) = 61	BCWS < 64.000	BCWS < 64.000	BCWS < 64.000	BCWS < 64.000	BCWP < 0.716
EAC(t) = 61	BCWS < 64.000	BCWS < 64.000	BCWS < 64.000	SPI < 94.500	BCWP < 0.716
EAC(t) = 61	BCWS < 64.000	BCWS < 64.000	SPI < 94.500	BCWP < 0.716	
EAC(t) = 61	BCWS < 64.000	BCWS < 64.000	SPI < 94.500	BCWS < 64.000	BCWP < 0.716
EAC(t) = 61	BCWS < 64.000	BCWS < 64.000	SPI < 94.500	SPI < 94.500	BCWP < 0.716
EAC(t) = 61	BCWS < 64.000	SPI < 94.500	BCWP < 0.716		
EAC(t) = 61	BCWS < 64.000	SPI < 94.500	BCWS < 64.000	BCWP < 0.716	
EAC(t) = 61	BCWS < 64.000	SPI < 94.500	BCWS < 64.000	BCWS < 64.000	BCWP < 0.716
EAC(t) = 61	BCWS < 64.000	SPI < 94.500	BCWS < 64.000	SPI < 94.500	BCWP < 0.716
EAC(t) = 61	BCWS < 64.000	SPI < 94.500	SPI < 94.500	BCWP < 0.716	
EAC(t) = 61	BCWS < 64.000	SPI < 94.500	SPI < 94.500	BCWS < 64.000	BCWP < 0.716
EAC(t) = 61	BCWS < 64.000	SPI < 94.500	SPI < 94.500	SPI < 94.500	BCWP < 0.716

Fig. 5 Association Rule results

Fig. 5 shows some rules in *IF ANTECEDENT THEN CONCEQUENT* forms. The support and confidence of these rules are more than minimum support and confidence which is determined by decision maker that is 10% and 30%, respectively, in our study. This information is beneficial to decide about the forecasting method for EAC (t). For example it could be seen that EAC (t) will be around 60 days when SPI is more than 0.94. Therefore, this typical analysis could be extended to achieve a reliable forecasting for completion time of a project.

IV. Conclusion

Determining the time and cost of a project during the project execution is very important. EVMS is a common methodology to measuring time and cost during the project execution. In this paper we have represented an application of data mining tools to predicate duration of a project. There are 6 tools in literature for data mining that we have used three of them here, for the nature of our problem. Because of proper forecasting by Data mining, project management team could plan preventive actions and these results would be applicable in practice. It could be seen that each factor have an specific role in estimate but in generalizing final decisions one should be aware of selection best parameters to change them to improve the project progress. Finally, it is obvious that a lot of researches could be followed by using data mining approach and this study is only one evidence for this purpose.

References

[1] D.F. Cioffi, 2006, “Designing project management: A scientific notation and an improved formalism for earned value calculations.” International journal of project management. Vol.24, no.2, 136-144.

[2] D.F. Cioffi, 2005, “A tool for managing project: an analytical parameterization of the S-curve”, international journal of project management, vol.23, no.3, 215-222.

[3] F. Anbari, 2003, “Earned value project management method and extensions”, Project management journal, vol.34, no.4, 12-23.

[4] G.Vinter, S. Rozenes, S. Spraggett, 2006, “Using data envelope analysis to compare project efficiency in a multi-project environment” International Journal of project management. Vol.24, no.4, 323-329.

[5] H.Iranmanesh, N.Mojir, S. Kimiagari.2007 “A new formula to estimate at completion of a project’s time to improve earned value management system”. IEEM, Singapore.

[6] Q.W. Fleming, J.M. Koppelman. 2000, “Earned value project management”. 2nd ed. Newtown Square, NJ: project Management Institute, INC.

[7] Vanhoucke, Vandevoorde, 2005.”A simulation and evaluation of earned value metrics to forecast the project duration”. Faculteit Economic En Bedrijfskunde.

[8] D.C. Bower. 2007. “New Directions in Project Performance and Progress Evaluation”. A thesis submitted to fulfil the requirements for the Degree of Doctor of Project Management. School of Construction, Property and Project Management RMIT University Melbourne, Australia

[9] D.T. Larose, “Discovery Knowledge In Data: an introduction to data mining”, Published by John Wiley & Sons, Inc., Hoboken, New Jersey, 2005.

Appendix

Sample Data of the Study

Period	CPI	SPI	BCWS	BCWP	ACWP	AD	PD	ED	EAC(t)
1	0.955	0.970	24	23	24	76	64	74	66
2	0.969	0.920	93	85	88	76	64	70	69
3	0.993	0.860	134	115	116	76	64	65	74
4	1.080	0.821	196	161	149	76	64	62	78
5	1.111	0.790	261	206	186	76	64	60	81
6	1.064	0.738	348	257	241	76	64	56	86
7	1.020	0.749	428	320	314	76	64	57	85
8	1.056	0.712	578	412	390	76	64	54	89
9	1.057	0.784	699	548	518	76	64	60	81
10	1.049	0.718	913	655	625	76	64	55	89
11	1.031	0.678	1162	788	764	76	64	52	94
12	1.009	0.701	1359	952	944	76	64	53	91
13	1.023	0.714	1519	1085	1061	76	64	54	89
14	0.999	0.721	1684	1214	1215	76	64	55	89
15	0.972	0.719	1899	1365	1405	76	64	55	89
16	0.943	0.728	2103	1531	1624	76	64	55	89
17	0.944	0.738	2358	1740	1843	76	64	56	87
18	0.949	0.723	2645	1912	2016	76	64	55	89
19	0.947	0.718	2948	2116	2235	76	64	55	90
20	0.951	0.698	3349	2339	2459	76	64	53	92
21	0.935	0.695	3679	2557	2735	76	64	53	93
22	0.920	0.717	3948	2831	3077	76	64	54	90
23	0.909	0.733	4168	3057	3363	76	64	56	88
24	0.918	0.740	4404	3261	3554	76	64	56	87
25	0.907	0.750	4586	3440	3792	76	64	57	86
26	0.889	0.760	4790	3642	4099	76	64	58	85
27	0.864	0.770	5033	3875	4485	76	64	59	84
28	0.843	0.772	5392	4164	4939	76	64	59	84
29	0.852	0.757	5771	4369	5128	76	64	58	86
30	0.853	0.737	6162	4540	5320	76	64	56	89
31	0.849	0.716	6586	4719	5557	76	64	54	92
32	0.837	0.709	6951	4927	5886	76	64	54	93
33	0.832	0.723	7196	5206	6260	76	64	55	91
34	0.822	0.739	7483	5531	6725	76	64	56	89
35	0.819	0.746	7866	5864	7159	76	64	57	88
36	0.824	0.745	8247	6143	7459	76	64	57	88
37	0.821	0.747	8555	6393	7784	76	64	57	88
38	0.821	0.755	8871	6698	8158	76	64	57	87
39	0.806	0.762	9099	6934	8597	76	64	58	86
40	0.799	0.782	9277	7259	9091	76	64	59	83
41	0.800	0.803	9474	7609	9514	76	64	61	81
42	0.798	0.813	9658	7848	9834	76	64	62	79
43	0.792	0.823	9811	8077	10204	76	64	63	78

44	0.791	0.834	9985	8324	10527	76	64	63	77
45	0.781	0.843	10174	8580	10989	76	64	64	76
46	0.773	0.851	10333	8797	11387	76	64	65	75
47	0.761	0.868	10429	9057	11909	76	64	66	73
48	0.761	0.892	10522	9387	12333	76	64	68	70
49	0.762	0.891	10647	9485	12443	76	64	68	71
50	0.757	0.888	10806	9596	12681	76	64	67	71
51	0.747	0.890	10977	9773	13076	76	64	68	70
52	0.750	0.896	11058	9909	13212	76	64	68	70
53	0.745	0.901	11092	9995	13408	76	64	68	69
54	0.744	0.906	11136	10091	13558	76	64	69	69
55	0.742	0.907	11207	10169	13713	76	64	69	69
56	0.738	0.912	11250	10256	13903	76	64	69	68
57	0.734	0.918	11293	10363	14128	76	64	70	67
58	0.735	0.924	11325	10465	14248	76	64	70	67
59	0.733	0.928	11375	10552	14396	76	64	71	66
60	0.729	0.930	11442	10645	14603	76	64	71	66
61	0.726	0.935	11498	10747	14803	76	64	71	66
62	0.722	0.941	11554	10873	15065	76	64	72	65
63	0.715	0.955	11566	11047	15445	76	64	73	63
64	0.715	0.960	11606	11140	15586	76	64	73	63
65	0.715	0.956	11669	11158	15613	76	64	73	63
66	0.715	0.958	11669	11183	15639	76	64	73	63
67	0.714	0.961	11669	11208	15698	76	64	73	63
68	0.713	0.965	11669	11256	15778	76	64	73	62
69	0.711	0.970	11669	11315	15921	76	64	74	62
70	0.709	0.974	11669	11368	16028	76	64	74	61
71	0.706	0.979	11669	11428	16182	76	64	74	61
72	0.702	0.986	11669	11509	16392	76	64	75	60
73	0.703	0.990	11669	11549	16438	76	64	75	60
74	0.700	0.993	11669	11592	16560	76	64	75	59
75	0.696	0.999	11669	11658	16739	76	64	76	59
76	0.697	1.000	11669	11669	16739	76	64	76	59