IIASA
International Institute for
Applied Systems Analysis
www.iiasa.ac.at

Distributed Performance Analysis on the Internet Using a Centric Database

Miettinen, K.M. and Korhonen, P.

IIASA Interim Report
December 1998

IIASA

International Institute for Applied Systems Analysis • A-2361 Laxenburg • Austria
Tel: +43 2236 807 • Fax: +43 2236 71313 • E-mail: info@iiasa.ac.at • Web: www.iiasa.ac.at

INTERIM REPORT IR-98-111/December

Distributed Performance Analysis on the Internet Using a Centric Database

Kaisa Miettinen(miettine@mit.jyu.fi)
Pekka Korhonen(korhonen@iiasa.ac.at)

Approved by

Gordon MacDonald(macdon@iiasa.ac.at)
Director, IIASA

Interim Reports on work of the International Institute for Applied Systems Analysis receive only limited review. Views or opinions expressed herein do not necessarily represent those of the Institute, its National Member Organizations, or other organizations supporting the work.

Contents

1. Introduction 6

2. Development of the Approach 7

2.1. Description of the System 7

2.2. Technical Realization 9

2.3. Regression Analysis Model 10

3. The Use of the Approach 11

4. Discussion 14

5. Conclusions 14

References 15

Abstract

In many areas of life it is useful to be able to compare one's own performance to some general benchmark data. The Internet provides a way for realizing such a comparison so that the original database can be hidden from users by locating it in a server computer and users can test their individual data in a distributed manner. An interactive and graphical user interface can be implemented with the tools of the World-Wide Web (WWW).

We introduce a World-wide INTEractive Regression Analysis (WINTERA) system that operates via the Internet. The system enables a user to carry out regression analysis with an original database and evaluate the performance of the data vector of her or his own. There are two kinds of users in the system. Data suppliers enter their observation matrices to form databases. Ordinary users can evaluate observations of their own with respect to the existing databases. They can also suggest their observations to be included in the databases. The data supplier decides whether (s)he accepts or rejects the information. This means that the whole database is accessible only to the data supplier. In any case, ordinary users receive information about their performance.

Keywords: Performance Analysis, Internet, WWW, Regression Analysis, Centric Database

Acknowledgments

This research was partly supported by the Academy of Finland, grant #22346. The authors wish to thank Mr. Tapani Tarvainen for his contribution to the implementation.

About the Authors

Pekka Korhonen was Project Leader of the Decision Analysis and Support Project (DAS) at IIASA during the period August 1997 - December 1998. He has returned to the Helsinki School of Economics and Business Administration where he is Professor of Economics and Business Administration.

Kaisa Miettinen was a research scholar with the DAS Project during the period September - December 1998. She returned to the University of Jyväskylä at the end of 1998 to continue her work as a researcher in the Department of Mathematical Information Technology.

Distributed Performance Analysis on the Internet Using a Centric Database

Kaisa Miettinen
Pekka Korhonen

1. Introduction

In many different areas of life it is often desirable to compare one's own performance to some general benchmark data. Examples of such areas where tools of comparative analysis are needed are the evaluation of the (financial) performance of firms, the evaluation of (high-technology) products and self-evaluation of bank customers.

The comparison can be supported with computer programs. Let us mention as an example a simple system developed at the University of Jyväskylä where students can evaluate their possibilities to enter the university. Another example is a system available at the web site of a Finnish newspaper where people may estimate their lifetime as the function of their inheritable characteristics and living habits.

The benchmark data may contain confidential information (like business secrets or personal information). In this case, the comparison of one's own performance is still possible: the details of the data are not shown to the user; only the results of the comparison and summary information are displayed.

Modern computer facilities provide new tools for realizing performance analysis systems. Of particular interest is the rapidly spreading use of the Internet with its graphical user interface World-Wide Web (WWW). Because the Internet is easily accessible, software on the Internet is automatically available to large numbers of people. If, in addition, the software operates via the Internet, no special requirements are set on the software or the operating system in the user's computer. Almost anybody can use the system. The main issue is that the only requirements are a connection to the Internet and a WWW browser. Furthermore, the WWW enables a convenient and graphical user interface with visualization possibilities.

The Internet enables centralizing the data to one computer and distributing the user interface to the computers of each individual user. In other words, we may have a centric database located in a server computer so that all the analysis takes place in the server. This means data security for concealed information. In addition to the data security, the realization of a performance analysis system via the Internet has several other advantages. The centric database is easy to update and maintain because there is only one version of the system and it is always the latest version that is available for every user. No updates of the databases or the analysis software have to be distributed. Thus, the Internet provides an excellent, flexible environment to realize our comparison approach.

Commercial services can also be established based on this idea of a centric database and distributed interface. If someone has a database and an analysis software, (s)he can easily offer performance analysis services via the Internet.

Here we introduce a World-wide INTEractive Regression Analysis system called WINTERA. As its name suggests, the performance analysis is based on interactive multivariate regression analysis. The system operates on the Internet and is available to everyone.

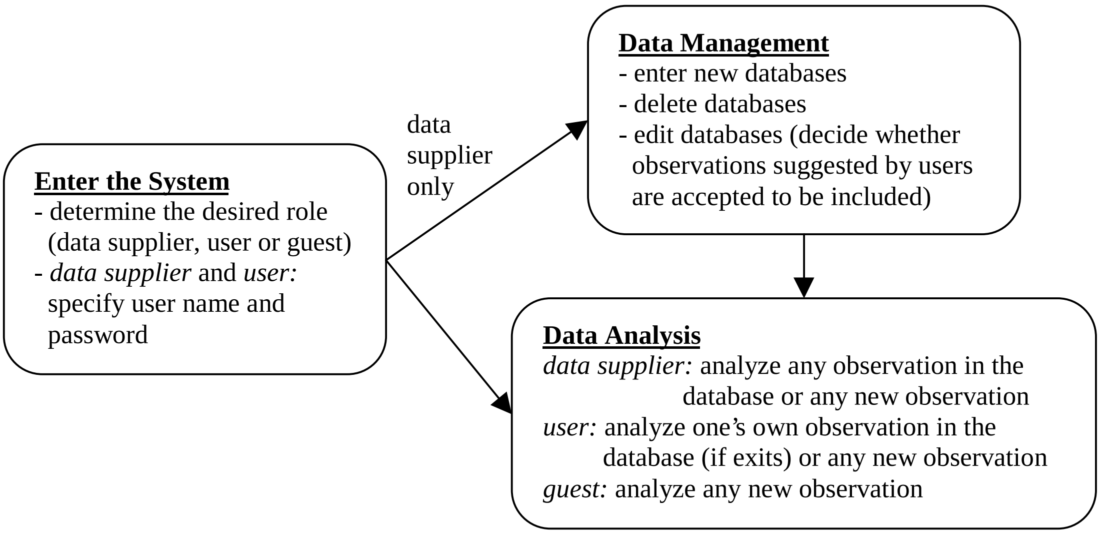
In the next sections we describe how WINTERA has been developed and realized. Then we briefly consider computational aspects in dealing with a regression model. We illustrate the functioning of the WINTERA system with a numerical example and conclude the paper with some future directions.

2. Development of the Approach

Among one of the starting points behind the development of the WINTERA system has been the goal of providing a versatile and flexible regression analysis tool with an interactive user interface. Another further principle has been not to allow ordinary users to have access to the underlying database. This is achieved by having two kinds of users in the system. In addition to ordinary users we have people who provide and maintain databases. These people, called data suppliers, have access to the details of their own databases and can modify the data according to their wishes.

2.1. Description of the System

The main features of the WINTERA system from the point of view of its user are entering the system, data analysis and data management, as illustrated in Figure 1.



Enter the System
- determine the desired role (data supplier, user or guest)
- *data supplier* and *user:* specify user name and password

data supplier only

Data Management
- enter new databases
- delete databases
- edit databases (decide whether observations suggested by users are accepted to be included)

Data Analysis
data supplier: analyze any observation in the database or any new observation
user: analyze one's own observation in the database (if exits) or any new observation
guest: analyze any new observation

Figure 1. Main features of the system

When entering the system, new users must first themselves specify their user name and password, together with a desired role in the system. Personal user names enable saving and handling private data. (Guests can visit WINTERA without identifying themselves, but they cannot save any data to the system.) The role in the system means being either a data supplier or an ordinary user. Data suppliers are persons who provide databases in an observation matrix form to the system for all the users to utilize. The rows of an observation matrix are called observation vectors, the components of which represent the values of various variables characterizing observations. Ordinary users can evaluate their own performance with respect to the databases that data suppliers have provided. Data suppliers can carry out performance analysis as well.

Users may suggest their observation vector to be included in the database (maximum one vector per user in one database). However, it is the appropriate data supplier who finally decides whether the suggested observations are included or not. The ordinary users have access to their personal observations in the databases (if such observations exist). It is still an unsolved problem how to find incentives so that ordinary users provide proper observations to the system. That is why data suppliers have full authority in our system to accept or reject the observations provided by users.

Before any data analysis can be realized, the user (data supplier, ordinary user or guest) has to select the database to be used. The collection available consists of the databases that different data suppliers have entered in the system. In the actual data analysis, the user has to specify which observation vector's performance has to be analyzed. In addition, (s)he has to identify one dependent variable and optionally one or more independent variables.

If the user already has an observation in the selected database, these variable values are shown as default values to be analyzed. The data supplier may select any of the observations in the database to be analyzed. However, if the database is owned by another data supplier, the first-mentioned data supplier is treated as an ordinary user. All three types of users (users, data suppliers and guests) can specify their own observation vector whose performance is to be analyzed with respect to the underlying observation matrix.

There are a few options available for controlling the performance of regression analysis. The old and/or the new observation are/is optionally used in the parameter estimation of the model. A typical situation is that the old observation is used whereas the new one is not. The regression line can also be forced to pass through the origin.

Here it should be noted that the selected role of the user can be changed if necessary. A user becoming a data supplier simply means that (s)he can enter and maintain databases of her or his own. If a data supplier wants to become a user, all her or his databases are deleted from the system. At the same time, the observations of ordinary users included in these databases are deleted.

Data management is only allowed for data suppliers and only as far as their own databases are concerned. In other words, each database has an owner who is the only person with access to details of the data. Associated to each observation vector in each database is information about its owner. Usually most of the observations are owned by the data supplier of that particular database. Each user may have a maximum of one observation in each database. The data supplier decides whether the observations suggested by users are to be included in the database.

(S)he can also edit the data, enter whole new databases or delete databases. If a database is deleted, all its observations specified by ordinary users are deleted as well.

2.2. Technical Realization

WINTERA is a program consisting of a user interface (in Python) and an underlying regression analysis code. The analysis software can be coded in any language. We used Pascal.

The master module of the system is the interface which transfers information between the user and the system. Furthermore, it is responsible for controlling the database and calling the regression analysis software with appropriate parameter values.

The user interface of the WINTERA system is a set of WWW pages created dynamically during the session. Thus, the user interface is a program creating pages described in HTML (hypertext markup language) and transferring information between the client and the server. The role of a browser is to interpret the HTML code. The browser of the user is a client of the server computer. The server computer is responsible for data management and computation whereas the client takes care of input and output.

WINTERA has been implemented by using standard programming facilities and uncommercial tools in order to keep the system as general as possible. This means that special features available only in certain browsers have been avoided. When considering secret data, a Java-type [1] distributed way of thinking is not reasonable. Instead, databases are centralized into one efficient server computer and saved there.

The programs creating HTML pages are called Common Gateway Interface (CGI) programs. We have implemented the interface with the object-oriented Python language [5] for its capabilities of dynamical handling of strings and the existing CGI modules for employing environmental variables. We have also been able to utilize some parts of an existing NIMBUS system [2, 4] for nonlinear multiobjective optimization. To our knowledge, NIMBUS is the first interactive multiobjective optimization software operating via the Internet. It is also based on the idea of centralized computing and distributed interface. (Quite recently, another web-based software called Web-HIPRE [8] has appeared for decision analytic problem structuring, multicriteria evaluation and prioritization.)

Besides the separate help pages, the homepage of WINTERA is the only fixed page. All the other twelve pages of the system are generated dynamically. The connections of the pages are depicted in Figure 2. From each page generated by the system the user may select how to continue. The next page is generated according to this information. For example, data suppliers have more choices because they can both control databases and analyze observations.

The password, contact information or usage role can be changed on the user information page that can be reached from almost every page. For clarity, none of those arrows are depicted in Figure 2.

The WINTERA system was first written as an implementation-independent pseudo-code. This is good practice for sketching the different features of the system and interconnections between its different properties. The pseudo-code also facilitates the documentation part.

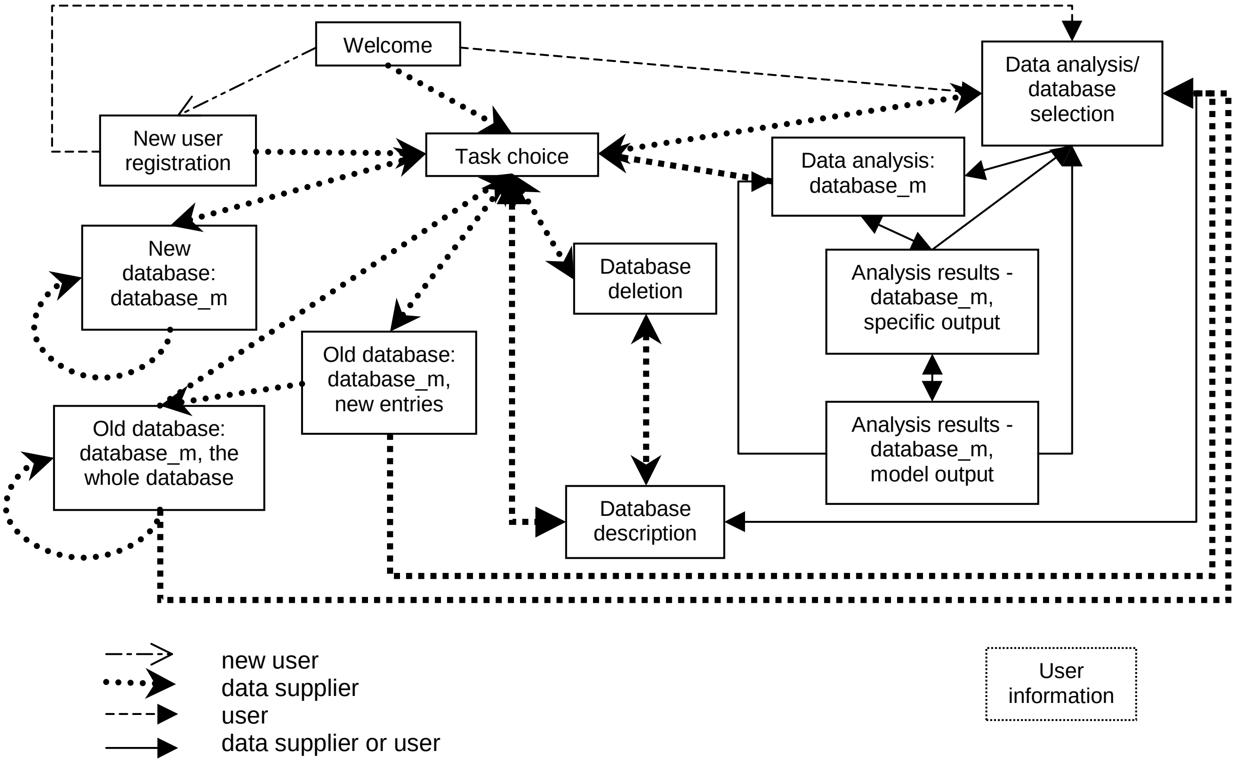

Welcome
Data analysis/ database selection
New user registration
Task choice
Data analysis: database_m
New database: database_m
Database deletion
Analysis results - database_m, specific output
Old database: database_m, new entries
Old database: database_m, the whole database
Analysis results - database_m, model output
Database description
new user
data supplier
user
data supplier or user
User information

Figure 2. Connections of different pages

2.3. Regression Analysis Model

Our tool in evaluating performance is multivariate regression analysis (see, for example, [3]). The database is given as an $n \times p$-observation matrix X, where n refers to the total number of observations and p to the number of variables. In the current version, the number of observations is not restricted in any way but the upper bound for the number of variables is 30. Each observation has a value for the p variables. Typically, a database has one or several performance variables and a set of potential explanatory (background) variables. The performance variables are used as dependent variables and the explanatory variables are independent variables. The role of the variables is specified while using the system.

Each observation in the database has an 'owner.' The owner can be either the data supplier or an ordinary user (each ordinary user may only have one observation in the matrix). The user of the WINTERA system may either analyze the performance of her or his observation existing in the database (if so) or that of a new observation. The corresponding observation vectors are here denoted by $\mathbf{x}_0$ and $\mathbf{x}_1$, respectively.

With some options the user can control how the regression analysis is carried out. First (s)he must specify one dependent variable and k independent variables ($0 \leq k < p$). One can also determine whether the old values (vector $\mathbf{x}_0$) and/or the new values (vector $\mathbf{x}_1$) are used in the parameeter estimation. Furthermore, the regression line can be forced to pass through the origin.

Based on the observation matrix X we generate an augmented cross-product matrix M to be called a model matrix

$$M = \begin{pmatrix} X^T X & X^T \mathbf{1} & \mathbf{x}_0 & \mathbf{x}_1 \\ \mathbf{1}^T X & n & 1 & 1 \\ \mathbf{x}_0^T & 1 & 1 & 0 \\ \mathbf{x}_1^T & 1 & 0 & -1 \end{pmatrix},$$

where $\mathbf{1}=(1,1, \dots , 1)^T$, $\mathbf{x}_0$ is an old observation vector (if available) and $\mathbf{x}_1$ is a new observation vector.

Note that the matrix *M* contains no detailed information about the other observations except the user's own one. Thus, the matrix *M* can be handled without losing any confidentiality of the database.

The matrix *M* contains all the necessary information needed for regression analysis. The results needed are obtained by applying various diagonal pivot operations to the model matrix (as described in [6-7]). This is a computationally efficient way for realizing the interactive analysis. Changes in the roles of the variables, observation values or in the roles of the observations (to be used in the parameter estimation or not) can be taken into account by applying pivot operations to the previously calculated model matrix. We have no need to operate with the original observation matrix during the analysis.

The output consists of two parts: specific output and model output. In the specific output, we provide information about the performance of the dependent variable. In addition to the current value, we show its explained value $\hat{y}$ at the current level of independent variables and calculate the probability of having a higher value than the current one. The result is also interpreted verbally. In the model output, we provide the standard output of regression analysis: R^2, *R* and adjusted *R*, *F*-values, regression coefficients, their standard deviations, etc.

3. The Use of the Approach

Here we demonstrate the functioning of the WINTERA system with a simple example. Data analysis can be carried out after entering the system and after selecting the appropriate database.

Our observation matrix contains a number of 26 observations. They describe different firms in the renovation business located in different places. The variables are renovation work carried out in a year (in hours), promotional expenditures (in ten thousand dollars), number of clients, number of competitors and a district potential (coded).

The whole database is given in Table 1. Note that ordinary users cannot see this information in the actual system.

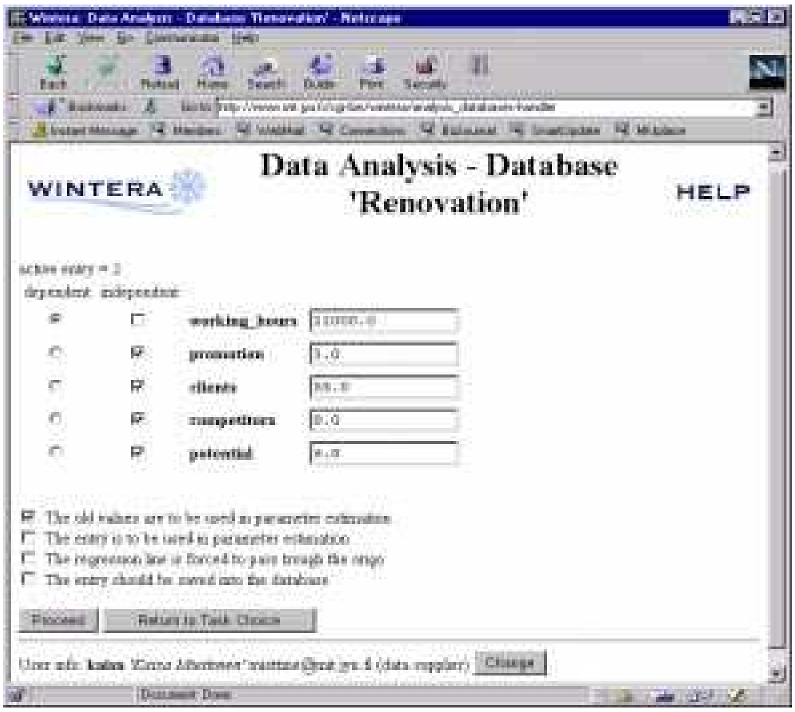
A data supplier can examine the whole database and edit it. (S)he can also select one observation as an active one whose components appear as default values in the data analysis. (If a user has an observation in the database, it is the default one. Otherwise, no default values are given by the system.)

Table 1. The observations

working hours	promotion	clients	competitors	potential
7930	5.5	31	10	8
20010	2.5	55	8	6
16320	8	67	12	9
20010	3	50	7	16
14600	3	38	8	15
17770	2.9	71	12	17
3090	8	30	12	8
29190	9	56	5	10
16000	4	42	8	4
33940	6.5	73	5	16
15960	5.5	60	11	7
8630	5	44	12	12
23750	6	50	6	6
10720	5	39	10	4
15500	3.5	55	10	4
29140	8	70	6	14
10020	6	40	11	6
13580	4	50	11	8
22330	7.5	62	9	13
19500	7	59	9	11
7340	6.7	53	13	5
4770	6.1	38	13	10
14070	3.6	43	9	17
9350	4.2	26	8	3
25900	4.5	75	8	19
33120	5.6	71	4	9

The database provides an entrepreneur with the possibility to test her or his alternative plans to develop the business. In our example, we assume that the firm represented by the second observation (20010.0, 2.5, 55.0, 8.0, 6.0) is to be re-evaluated. The owner of the firm is willing to expand the business. (S)he thinks that by increasing the promotional expenditures by 5000 dollars the number of working hours can be increased up to 22.000 hours. Thus, the new observation vector is (22000.0, 3.0, 55.0, 8.0, 6.0).

The first variable (working hours) is the dependent variable describing the performance of the business, and all the others are treated as independent variables. The data analysis page of WINTERA is given in Figure 3. The old observation vector is used in the parameter estimation whereas the regression line is not forced to pass through the origin.



Data Analysis - Database 'Renovation'

active entry = 2

dependent independent

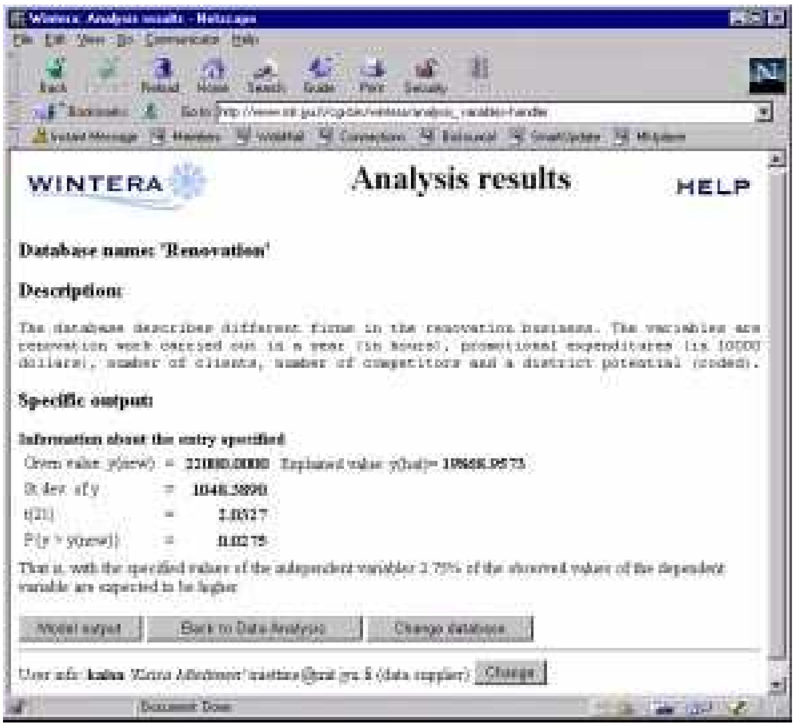
●	□	**working_hours**	22000.0
○	☑	**promotion**	3.0
○	☑	**clients**	55.0
○	☑	**competitors**	8.0
○	☑	**potential**	4.0

☑ The old values are to be used in parameter estimation
□ The entry is to be used in parameter estimation
□ The regression line is forced to pass trough the origo
□ The entry should be stored into the database

User info: **kaisa** 'Kirsi [illegible]' [illegible] (data supplier)

Figure 3. The data analysis page of WINTERA

The regression analysis results consist of two parts: specific output and model output. The specific output (in Figure 4) describes the performance of the specified observation.



Analysis results

Database name: 'Renovation'

Description:

The database describes different firms in the renovation business. The variables are renovation work carried out in a year (in hours), promotional expenditures (in 10000 dollars), number of clients, number of competitors and a district potential (coded).

Specific output:

Information about the entry specified

```
Given value  y(new)  =  22000.0000   Explained value  y(hat)= 19868.9573
St. dev. of y        =   1048.3890
t(21)                =      2.0327
P(y > y(new))        =      0.0275
```

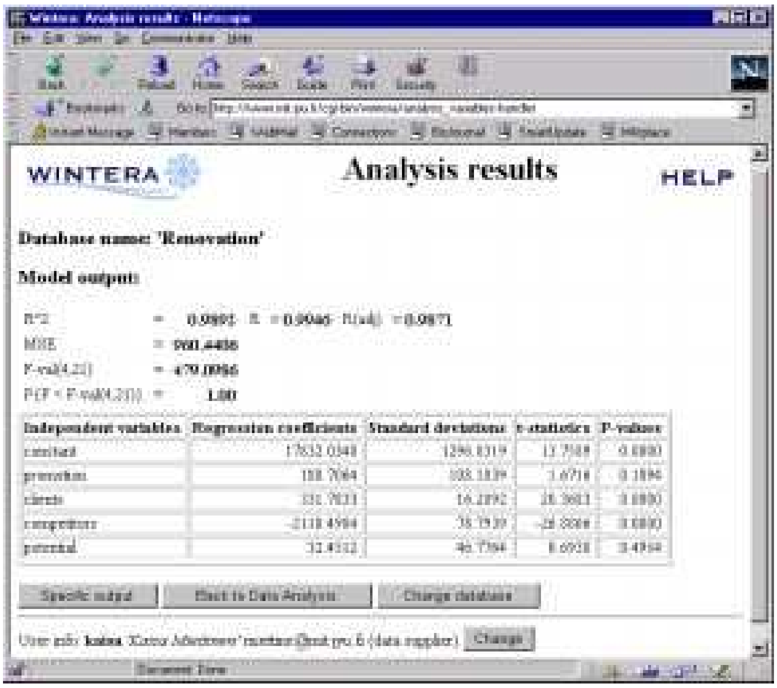
That is, with the specified values of the independent variables 2.75% of the observed values of the dependent variable are expected to be higher

User info: **kaisa** 'Kirsi [illegible]' [illegible] (data supplier)

Figure 4. Specific output of WINTERA

In our example, the explained value of the variable working hours is 19869.0 whereas the given value was 22000.0. With the specified values of the independent variables (promotion, clients, competitors and potential), 2.8% of the observed values of the dependent variable are expected to be higher. Thus, the idea of expanding the business in the above-mentioned way was rather unrealistic.

The model output (in Figure 5) contains standard regression analysis results.



Analysis results

Database name: 'Renovation'

Model output:

```
R^2               =   0.9891  R  = 0.9946  R(adj)  = 0.9871
MSE               =  980.4456
F-val(4,21)       =  479.0956
P(F < F-val(4,21)) =      1.00
```

Independent variables	Regression coefficients	Standard deviations	t-statistics	P-values
constant	17832.0348	1398.0319	12.7589	0.0000
promotion	118.7064	108.1839	1.0718	0.1896
clients	331.7823	16.2092	20.3603	0.0000
competitors	-2118.4984	78.7939	-26.8884	0.0000
potential	32.4512	46.7394	0.6938	0.4954

User info: **[illegible]** [illegible] (data supplier)

Figure 5. Model output of WINTERA

4. Discussion

We have emphasized that a centric database is safe if the observation matrices contain confidential information. It is true that in our implementation the users of the system have no access to the underlying data. However, the system administrator of the server computer naturally has access to it. If this is to be avoided, the observation information must be crypted as it is inputted in the system and this crypting must be opened when the model matrix is to be formed. The original observation matrix cannot be deduced from the model matrix, thus no crypting is necessary after this phase.

5. Conclusions

We have demonstrated how performance analysis software can be implemented with the help of the Internet and its graphical user interface WWW. As an example we have described an interactive system WINTERA for multivariate regression analysis. More developed performance analysis systems can be realized based on the same principles. It is easy to replace the underlying analysis tool and modify the interface where appropriate. Examples of possible further developments are applications of data envelopment analysis.

By centralizing the data in a server computer the users can only see the results of the analysis - not the original database. The advantage of an Internet implementation is also the wide accessibility of the system which almost anybody with any computer can use; all that is needed is a connection to the Internet and a WWW browser.

References

[1] Java Home Page. URL http://java.sun.com

[2] K. Miettinen and M.M. Mäkelä, Interactive MCDM Support System in the Internet, in: T.J. Stewart and R.C. van den Honert, Eds., Trends in Multicriteria Decision Making (Springer-Verlag, Berlin, Heidelberg, pp. 424-433, 1998).

[3] J. Neter, M.H. Kutner, C.J. Nachtsheim and W. Wasserman, Applied Linear Regression Models, 3rd Edition (Richard D. Irwin, Inc., Burr Ridge, Illinois, 1996).

[4] NIMBUS Home Page. URL http://nimbus.mit.jyu.fi

[5] Python Language Home Page. URL http://www.python.org

[6] H. Väliaho, A Synthetic Approach to Stepwise Regression Analysis, Commentationes Physico-Mathematicae 34, No. 12, pp. 91-132 (1969).

[7] H. Väliaho and T. Pekkonen, A Procedure for Stepwise Regression Analysis with a Program in FORTRAN V (Akademie-Verlag, Berlin, 1976).

[8] Web-HIPRE Home Page. URL http://www.hipre.hut.fi/