Building Simulation 2003

Eighth International IBPSA Conference
Eindhoven, Netherlands
August 11-14, 2003

APPLICATION OF DATA MINING TECHNIQUES FOR BUILDING SIMULATION PERFORMANCE PREDICTION ANALYSIS

Christoph Morbitzer [1], Paul Strachan [2] and Catherine Simpson [3]
[1] HLM Design, Glasgow, United Kingdom, email cmorbitzer@hlm.co.uk
[2] Energy Systems Research Unit (ESRU), University of Strathclyde, Glasgow, United Kingdom, email paul@esru.strath.ac.uk
[3] Building Simulation, Wiltshire, United Kingdom, email cathie@buildingsimulation.co.uk

ABSTRACT

Simulation exercises covering long periods (e.g.. annual simulations) can produce large quantities of data. The result data set is often primarily used to determine key performance parameters such as the frequency binning of internal temperatures. Efforts to obtain an understanding for reasons behind the predicted building performance are often only carried out to a limited extent and simulation is therefore not used to its full potential.

This paper describes how data mining can be used to enhance the analysis of results obtained from a simulation exercise. It identifies clustering as a particular useful analysis technique and illustrates its potential in enhancing the analysis of building simulation performance predictions.

INTRODUCTION

The amount of data generated from a simulation run can be considerable, depending on the number of days simulated. Different users of simulation programs have varying preferences for the duration of a simulation, varying from a typical day to annual simulations, depending on what is believed to be required to understand the behavior of the building. Many practitioners approach the assessment of a building by performing simulations that cover long periods, typically a year [Donn 1997], or even multi-year.

Figure 1 shows the results of an air flow analysis that was carried out for one zone of a simulation model over a two month period. It is straightforward to extract typical and extreme values for the air change rate in the zone, but specific questions are more difficult to answer, for example:

- Under what conditions does the air change rate in the building exceed 6 air changes per hour?
- How does wind speed and direction affect the air change rate in the zone?
- Under what conditions do comfort problems due to draughts occur ?

All of the above questions involve an analysis of several parameters (air change rate, air flow rate through an opening, wind speed, wind direction) which can change significantly during short periods. Patterns that give answers to the questions stated above are normally extracted by users viewing tabular or graphical data displays. Supporting this process with additional analysis techniques was identified as a useful contribution to enhance the capabilities of simulation and better integrate the tool into the building design process.

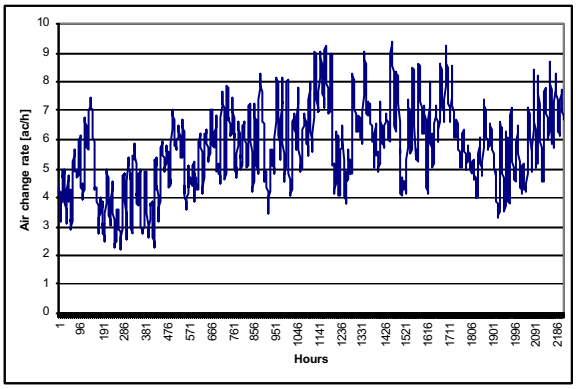

Figure 1 Display of air flow predictions obtained from a simulation exercise

POSSIBLE ANALYSIS TECHNIQUES

The search for knowledge (or patterns) in data is not a new concept, but was of interest even when data was stored in non-electronic form. Examples for pattern finding tools in electronic data sets that have been developed in the past are query functions of database management systems (DBMS). This section lists and rates techniques that could be applied for such an analysis (Table 1 summarises the techniques and their ratings).

One example for data analysis is its visual investigation, for example by displaying the data as a scatter plot graph. Figure 2 shows such a graph displaying heat extracted from a building by means of cooling versus the ambient temperature conditions.

Table1
Rating of different analysis techniques
(+ yes, - no, 0 neutral)

ANALYSIS TYPE	FAST AND INTERAC-TIVE ANALYSIS	NUMERICAL QUANTIFI-CATION OF FINDINGS	MULTIPLE VARIABLE ANALYSIS	EASY TO USE SOFTWARE IMPLEMEN-TATION	VISUALISA-TION OF FINDINGS
VISUAL ANALYSIS	+	-	0	+	+
REGRESSION ANALYSIS	+	+	0	-	0
UNCERTAINTY ANALYSIS	-	+	+	0	0
DATA MINING	+	+	+	+/0	+/0

Visual analysis is easy to carry out and can reveal useful information. However, the analysis does not provide numerical quantifications of findings and the analysis of a larger number of variables can be difficult.

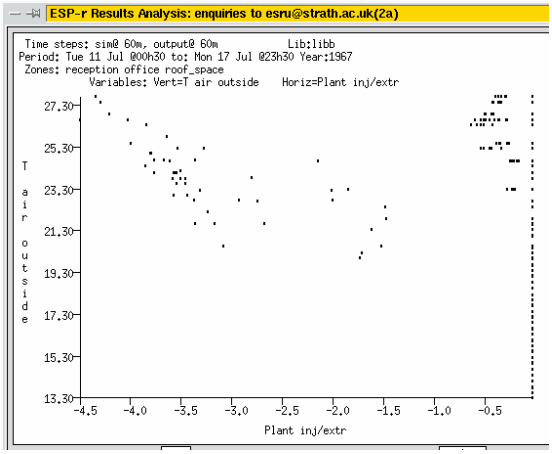

Figure 2 Scatter plot graph

Numerical analysis could be carried out with a regression analysis of the two variables. A linear least square analysis evaluates how far data deviates from its least-square line - the least-square line is the line that fits best the distribution of data points (see Figure 3). In the case of a strong linear correlation, the points lie close to the least-squares line and the sum of square distances between the points and their corresponding line values is small. Nonlinear regression analysis applies the same principal as a simple linear least square analysis but under the assumption that the variables have a non-linear correlation. Multiple regression analysis is an extension to simple linear regression analysis when more variables are added.

Regression analysis supports the analysis of data correlation by giving numerical information such as the correlation coefficient. Software tools which can be used to carry out the analysis are powerful but also complex and rather difficult to operate and often require from the user statistical background knowledge [Swain 2001, Thomas 1997, Lionheart Publishing 2001].

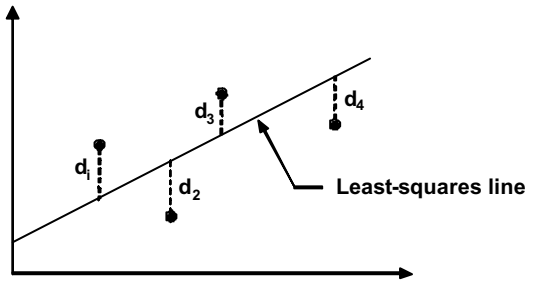

Figure 3 Least-squares fit

Uncertainty analysis techniques such as factorial analysis are another possibility for the evaluation of correlation between variables. Variable(s) under consideration are changed in automated, multiple simulation runs and the extent to which these changes affect the building performance is evaluated. By this means it is possible to determine the design parameters that have an impact on the behaviour of the building. Uncertainty analysis requires first the specification of the variables under consideration. After that the uncertainty is evaluated by means of multiple simulation runs. This process can be time and CPU intensive. Hence the approach does not allow the interactive analysis of a building design by focusing on different variables in turn within short time periods.

Data mining provides (at least to a certain extent) all of the requirements outlined in Table 1. It is possible to quickly and interactively analyse the data using existing generic data mining software. Many of the analysis processes are automated or at least semi-automated; hence the method allows the analysis to be carried out by a user with a very limited understanding of the underlying numerical analysis techniques. Findings are supported by both numerical quantification and visualisation and rules also often help the designer to understand patterns within the data set. The available software packages allow the analysis of multiple variables and can also evaluate categorical data. The description of data mining later in this paper will support these statements.

DATA MINING PROCESS

Figure 4 shows the different steps involved in the extraction of knowledge from data (after [Han and Kamber 2001]):

1. Cleaning of the data to remove noise or missing data[1].
2. Integrating the data into data warehouses – this is applied if multiple data sources are combined.
3. Selecting task relevant data.
4. Applying data mining to extract patterns from the data - here the user can choose between the different techniques which will be described later.
5. Evaluating the patterns that the data mining tool has discovered.
6. Presenting the significant patterns to the user.

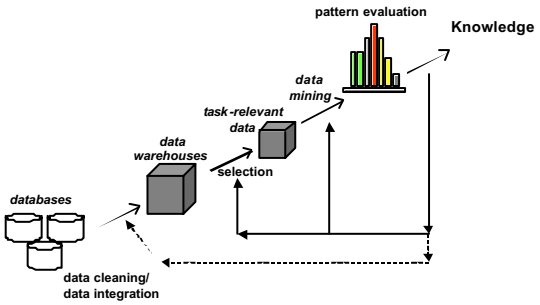

Figure 4 Data mining process

[1] This step is not really required when using data sets that have been generated from simulation programs. However, noise and missing data is an important aspect in the development of new data mining algorithms, because accurate data cleaning cannot be guaranteed and might cause "incorrect" knowledge.

Data mining does not automatically extract all available knowledge that is embodied in a data set. Although it may at sound at first appealing to have such an autonomous data mining system, in practice, such a system would uncover an overwhelmingly large set of patterns, and most of the patterns discovered in the analysis would be irrelevant for the user. A more realistic scenario is to communicate with the data mining system, using additional questions to examine the findings and direct the mining process (after [Han and Kamber 2001]):

- What is task relevant data?
- What kind of knowledge do I want to mine?
- What background knowledge could be useful?
- How do I want the discovered patterns to be presented?

In consequence the first analysis will not necessarily provide the required information– the user might have defined a mining exercise that does not reveal important patterns. In that case the analysis needs to be refined. The creation of different mining exercises is supported by a very flexible definition of a mining task. The user can quickly change variables to be included in a mining run, in combination with filters that can be defined for all the variables (e.g. only focus on times with a resultant temperature above 27°C, occupied periods, times of high occupancy densities etc.).

DATA MINING TECHNIQUES

This section describes and discusses different data mining techniques that could potentially used within the building design process to analyse performance predictions obtained from a simulation exercise.

Association Mining

Association mining discovers association rules that occur frequently together in a given set of data. Examples of association mining rules are:

wind speed (4-6) ∧ wind direction (270-300)

⇒ air change rate (4-6)

or

air change rate (4-6) ∧ resultant temperature (26-28) ⇒ solar radiation (200-400)

Association mining has a number of disadvantages: it produces a large number of redundant rules (see second of the above rules) and the visual displays

used to illustrate results from a simulation exercise can be difficult to comprehend (see Figure 5).

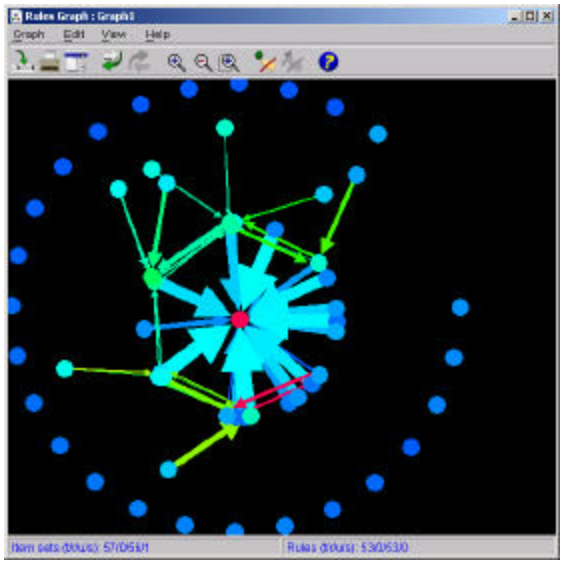

Figure 5 Association mining result display

(Tree) classification

Classification analysis intends to identify, as with the association mining technique, rules from the dataset it investigates. The difference is that association mining aims to discover any correlation between the different variables of the dataset, whereas classification mining only discovers rules that relate to one particular variable. In consequence, the user has to specify a "target" [Salford Systems 2000] or "active" [IBM 1999] variable in the process of setting up the mining task Example rules that could be obtained from the classification mining technique with air change rate as a target are:

wind speed (4-6) ∧ wind direction (270-300) ⇒ air change rate (4-6)

or

wind speed (2-4) ∧ wind direction (30-90) ⇒ air change rate (0-2)

with wind speed in [m/sec], wind direction in [°] and air change rate in [air changes/hour].

Classification mining results are often displayed in a tree format and are then referred to as 'tree classification'.

Tree classification has, in comparison with association mining, the advantage that it allows the definition of a mining focus by defining an active or target variable. However, the analysis of the mining results is rather tedious and visual display is (as for association mining) also different from display types normally used in the design process (see Figure 6).

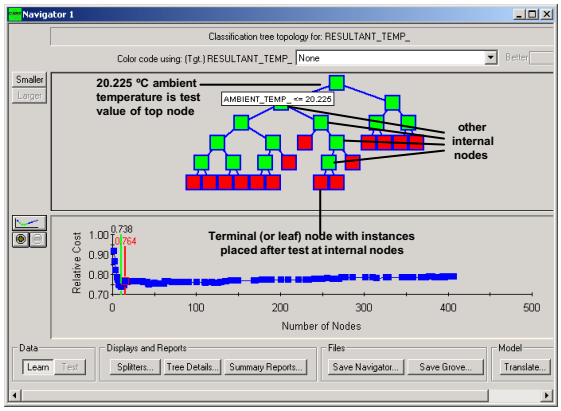

Figure 6 Classification mining result display

Outlier analysis

Normally data mining is used to extract typical patterns from a data set, e.g. under which conditions a room will overheat. However, a dataset may also contain instances (an instance is in this study a simulation hour) that do not comply with these typical patterns. These instances are called outliers. When using data mining in conjunction with simulation data analysis the designer is primarily interested in obtaining a general understanding of the behaviour of the building. Focus on exceptional patterns is only of secondary relevance (if at all). Outlier analysis is hence not discussed further.

Evolution analysis

Evolution analysis (or time series analysis) describes variables' behaviour over time and is an interesting concept that could be used for the evaluation of phenomena such as storage effects. However, it is rarely incorporated in data mining tools and has hence not been included in the evaluation.

Clustering

The output of a cluster analysis is different from the rules created by the association or classification mining technique. In a cluster analysis the data is grouped with the aim of placing instances in segments in a way that maximises the similarity between instances of one segment and minimises the similarity between the instances of different segments.

Cluster analysis was found to be the most suitable data mining technique for the analysis of simulation results data. The remainder of the paper focuses on this technique, including a case study.

CLUSTERING

Defining clustering mining parameters

With the cluster analysis data mining technique it is possible to carry out an analysis with continuous and categorical data sets (categorical analyse can be used for data sets with information such as different reference cases, design parameters, etc.). The user can also specify so-called active and supplementary variables. The active variables are used by the mining function when performing the clustering. The supplementary variables can be used to gain statistical information from the clusters that are found, determining the correlation between the cluster and supplementary variables.

Clustering results analysis

Figure 7 depicts the results obtained from a clustering exercise that analysed reasons for overheating in a room, focusing on periods with resultant temperatures above 27°C. The simulated room was 6.0 by 4.0 meters big, and 70% of one longer wall was fully glazed. This wall was south facing. The ventilation rate in the room was 2.0 achr in between 8:00h and 20:00h and 0.5 achr in other times. Resultant temperature was specified as the active variable and ambient temperature and direct solar radiation as supplementary variables. Note that in the display supplementary variables are indicated with rectangular brackets around the variable name.

Instances with similar characteristics have been grouped into *clusters* ([Agrawal et al 1998] describe the algorithms on which the analysis is based), as described above in this case in dependence of the resultant temperature.

The display shows six rows, each representing one of the clusters. Cluster 6, for example, contains data for simulated hours with resultant temperature just below 33°C, cluster 3 for resultant temperatures just above 27°C and the other clusters cover intermediate temperatures. The numbers down the left represent the cluster size as a percentage; for example, the top cluster contains 35% of the instances of the overall data set. The number on the right represents the cluster ID.

Figure 8 shows that each frequency binning chart of the cluster analysis display two different data sets. The solid bars represent the *data for the entire data set* and the transparent bars represent the distribution of the *instances that have been included into the particular cluster*.

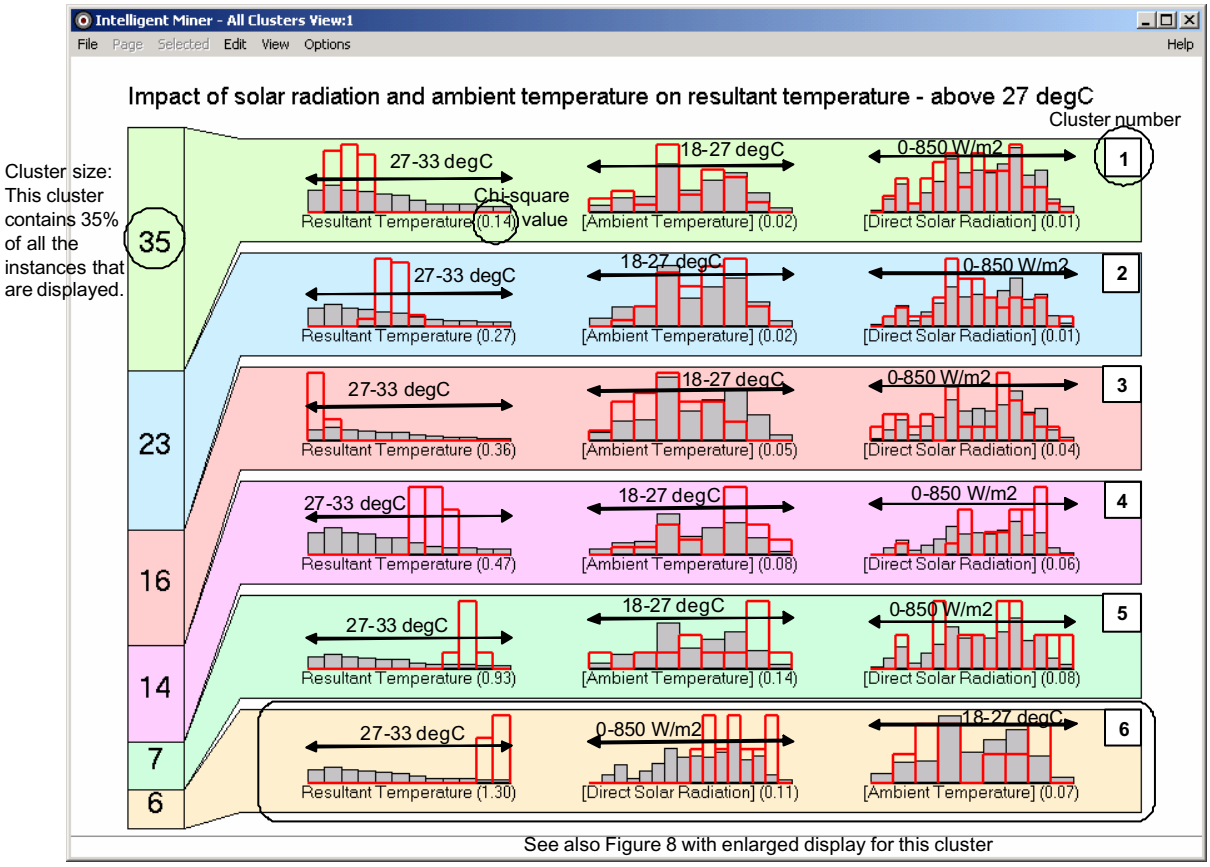

Figure 7 Cluster analysis results display

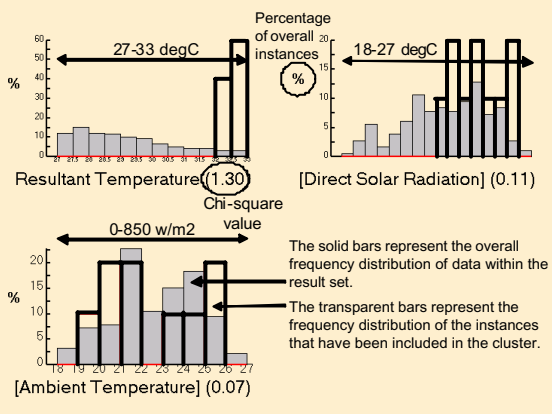

Figure 8 Display of single cluster

In addition to the clustering of the data the program has also established for every variable in each cluster the chi-square (or goodness of fit) value. This value is determined by comparing the frequency distribution of the entire data set with the frequency distribution of instances that have been included in the cluster (the solid and transparent bars). If the difference is small the chi-square value is small, and vice versa. Normally the active variable will have high chi-square values because the clusters were defined using this variable. If a supplementary variable also has a high chi-square value this indicates for this particular cluster a potential correlation between the two variables. The data mining program orders in its display variables for each cluster with respect to their chi-square value, with the variable with the highest chi-square value positioned on the left[2].

The interpretation can be illustrated with cluster 6 of the previous data mining exercise as depicted in Figure 8 From the analysis the designer can see that for this particular cluster, out of the two supplementary variables the direct solar radiation has the higher value, indicating a stronger correlation between high resultant temperature and direct solar radiation than between high resultant temperature and ambient temperature.

A general observation when viewing the entire results display in Figure 7 is that the chi-squared values are higher for clusters with high resultant temperatures. This can be explained as follows: average temperatures can occur under a number of different conditions, but extreme temperatures require particular conditions, which will result in stronger correlations and higher chi-squared values.

[2] This will normally be the active variable, but in exceptional cases it can also be a supplementary variable.

CASE STUDY

This case study illustrates how data mining can be used to support the designer in the planning of a natural ventilation scheme for a building. The building has a central atrium with adjacent office areas. Figure 9 displays the thermal and air flow model. It contained horizontal connections between the office zones and the atrium zones and vertical connections between the different zones specifying the atrium. For each office zone external openings were specified along the longer façade of the building and for the atrium two external openings were specified at the bottom of the shorter ends and one was specified at the top.

The simulation focused on the summer case and lasted from the 1st of May until the 30th of September. Figure 10 displays the air change rate in the atrium. Data mining was used to gain an initial understanding for the correlation between climatic conditions and air flow through the building

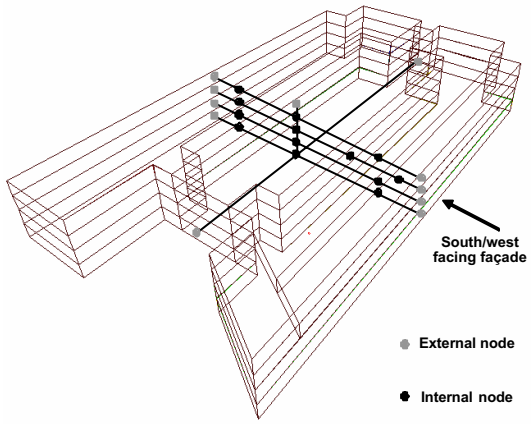

Figure 9 Simulation model used for case study

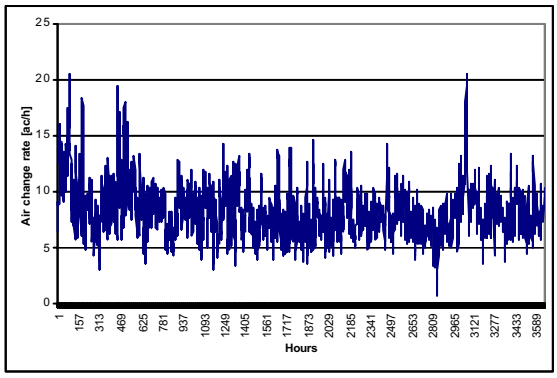

Figure 10 Predicted air change rate in atrium

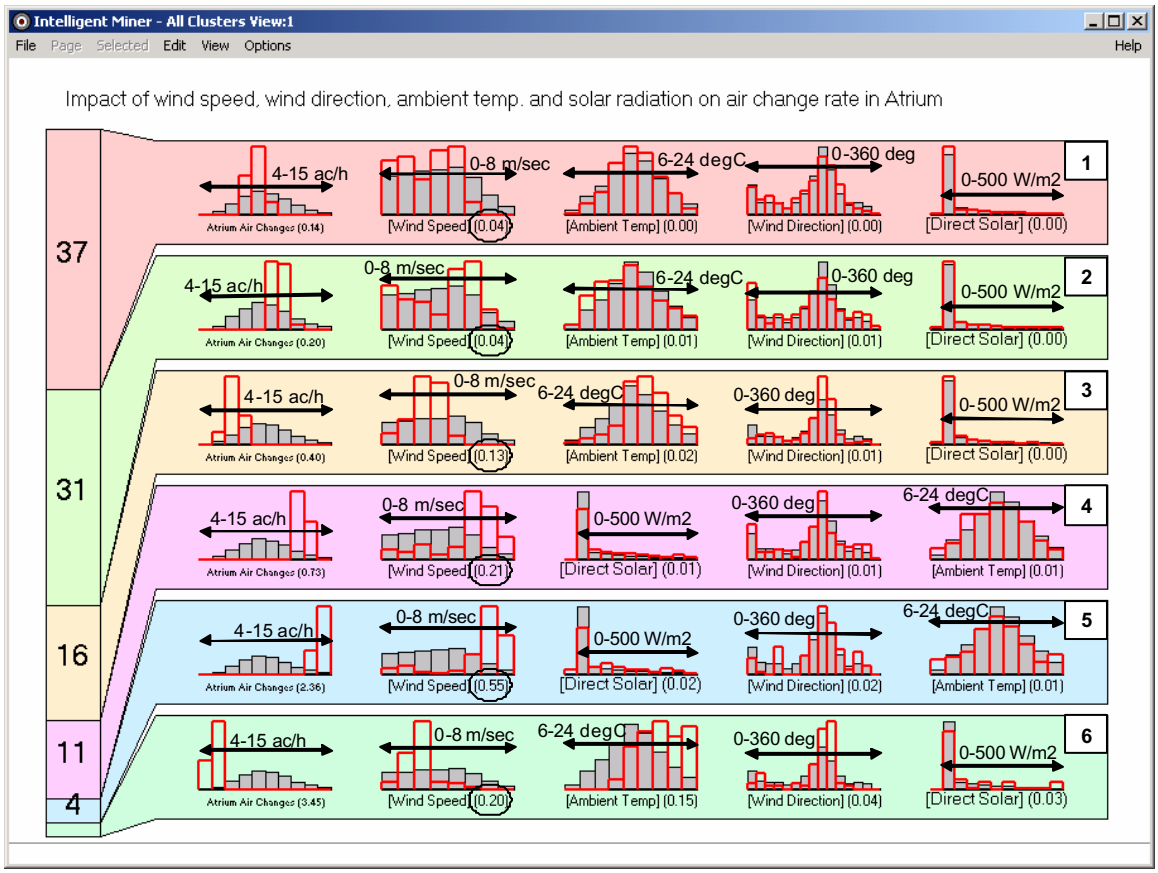

Figure 11 Cluster analysis results display

In a first data mining run the extent to which wind speed, wind direction, ambient temperatures and direct solar radiation influence the ventilation rate of the lower space in the atrium was evaluated. Figure 11 displays the outcome of an analysis using data mining. When examining the chi-square values of the different variables it can be seen that the highest correlation exists between wind speed and the air change rate in the space (see chi-square values for wind speed). This shows that wind speed is the most important driving force for air flow in the building.

In a second exercise, the influence of wind speed, wind direction, ambient temperature and direct solar radiation on the air flow directions within the building was examined. Figure 12 displays the results for the cluster with minimum upward or downward flow. It can be seen that this occurs in periods with wind speed above 3 m/s. This pattern was also confirmed by an independent analysis of numerous CFD and airflow network simulation results which showed that when the wind speed exceeds 3 m/s the wind pressure is so strong that air flow switches from upward ventilation to cross ventilation (see Figure 13).

Figure 14 shows another data mining analysis that underlines this finding. The figure displays air flow through the external opening of the south/west facing office space on the ground floor (positive data indicates air flow from the outside to the inside). It can be seen that air flow from the inside to the outside only occurs at times when the wind speed exceeds 3 m/s.

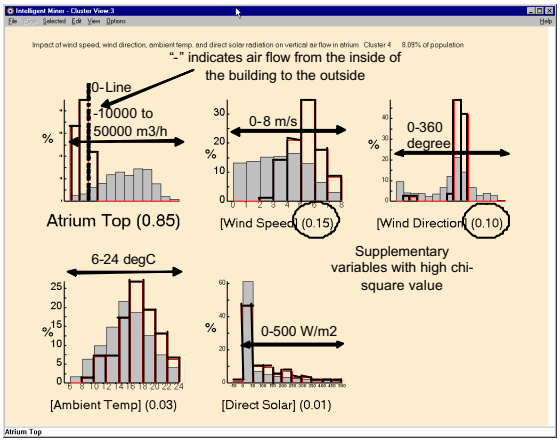

Figure 12 Small vertical air flow occurs with wind speed above 3 m/sec

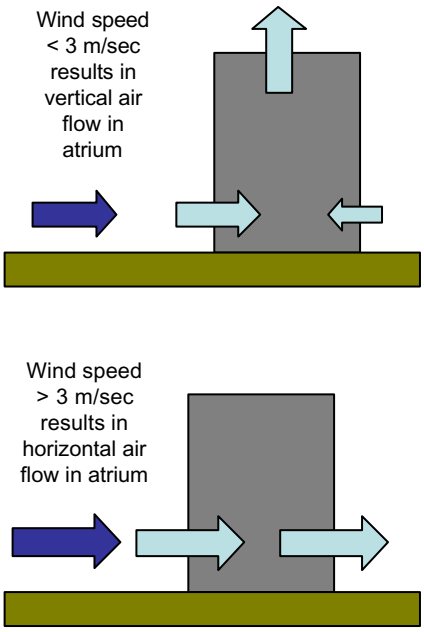

Figure 13: Change of air flow pattern with wind speed above 3 m/sec

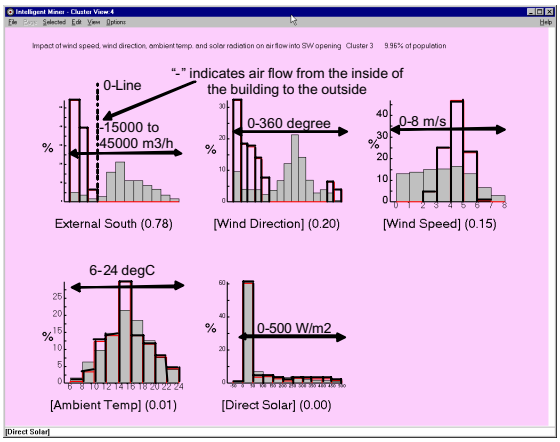

Figure 14: Cross ventilation occurs with wind speed above 3 m/sec

CONCLUSIONS

This paper described research into the application of data mining for the analysis of performance predictions obtained from a simulation exercise. The reason for research into the application of data mining was justified by the fact that the generation of large datasets with simulation requires data analysis with appropriate tools. Different data mining techniques were described and it was concluded that clustering was most applicable for this purpose. It had the most comprehensive visual display and in addition it indicated correlations between target and supplementary variables, independently for each cluster.

After that a case study illustrated the benefits of using data mining for the analysis of performance predictions obtained from a simulation exercise.

The case study also emphasized the particular usefulness of data mining for the analysis of combined thermal and air flow simulation, where boundary conditions constantly change and the system also has a quick response time.

The research has so far mainly focused on continuous data and on the evaluation of a single result set and not multiple simulation runs. These are potential areas for further investigation.

REFERENCES

Agrawal R, "Data Mining: Crossing the Chasm", Invited Talk at the 5th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-99), San Diego, California, August 1999.

Donn M R, "A Survey of Users of Thermal Simulation Programs", pp 65-72, Proceedings Building Simulation 97, Prague, pp 65-72, 1997.

Han J, Kamber M, "Data Mining – Concepts and Techniques", Morgan Kaufmann Publishers, 2001.

IBM, "Using the Intelligent Miner for Data", Version 6 Release 1, IBM, 1999.

Lionheart Publishing, "Statistical Analysis Software Survey", http://www.lionhrtpub.com/orms/surveys/sa/sa8.html, 2001 (viewed 2002).

Salford Systems, "Cart for Windows User's Guide", Salford System, 2000.

Swain J L, "Looking for Meaning in an Uncertain World – 2001 Survey of Statistical Analysis Software Products", http://www.lionhrtpub.com/orms/orms-10-01/survey.html , 2001 (viewed 2002).

Thomas L, Krebs C J, "A REVIEW OF Statistical Power Analysis Software", http://www.zoology.ubc.ca/~krebs/power.html, 1997 (viewed 2002).