Analyze building performance data for energy-efficient building operation

A. Ahmed, J. Ploennigs, Y. Gao & K. Menzel
IRUSE, University College Cork, Ireland

ABSTRACT: Modern buildings contain several sensors and meters to monitor the building performance. This data allows analyzing the building performance to increase the energy-efficiency along with user comfort. This paper presents two approaches to analyse building performance data. One solution uses data warehouse techniques to create sophisticated energy consumption aggregations. A second approach implements data mining techniques to estimate the thermal comfort of occupants with a reduced number of sensors. This paper interprets the knowledge gained using, as an example, University College Cork's Environmental Research Institute building to demonstrate the feasibility of this approach.

1 INTRODUCTION

There is a great interest to improve energy management in buildings considering the increasing price of fuel, and the global goal of reducing CO^2 output. *Building Energy Management* (BEM) aims at the effective and efficient usage of energy to maintain high building performance operation (Capehar et al. 2008, p. 1). One of the current challenges in this domain is to optimise energy consumption, while considering occupant comfort (Metz 2007, p. 394).

Building performance analysis emphasizes the measurement and assessment of various performance indicators covering the interests of owners, operators, and occupants in aspects like energy, lighting, thermal comfort, and maintenance (Augenbroe & Park 2005).

The continuous development of wired building automation systems and the current emerging of easy-to-integrate wireless solutions have increased the amount of available *building performance data* (Menzel et al. 2008) to evaluate these indicators. Traditional database management systems (DBMS) are nowadays used to store the building monitoring data. These DBMS lack the ability to create data aggregations and do not support the analysis of building performance data to deliver reports and actionable information (Lane 2007, p. 29).

Modern approaches from computer science may simplify the building performance analysis. *Data Warehouses (DW)* adds data aggregation capabilities to databases to prepare and deliver reports for large data sets (Stackowiak et al. 2007). They also facilitate the use of modern analysis approaches such as *Knowledge Discovery in Databases and Data Mining* (KDD) (Han & Kamber 2006, p. 35) to discover previously unknown characteristics, relationships, dependencies, or trends in data (Rob et al. 2008, p. 744).

The paper introduces a system that incorporates these two technologies to simplify the building performance analysis. Data Warehouse technologies are used to aggregate building performance data and provide to users a fast and easy way to manually analyse it. This approach is demonstrated in Section 2 for the energy consumption of a real building.

Data mining approaches can be used to analyse patterns in building performance data, but also to train models (Section 3). This is presented during the evaluation of thermal comfort to identify rooms with low comfort in Section 4 using only room temperature sensors. The data mining process is introduced from building data sources, to data preparation and transformation, model building, testing, and scoring.

The paper uses real data from the Environmental Research Institute (ERI 2002). The ERI is an energy-efficient building with many sustainable energy features such as solar panels, geothermal heat pumps and heat recovery systems. The ERI building is used by multiple research groups from biology, chemistry, as well as engineering. It also facilitates as a "Living Laboratory" to demonstrate smart building concepts. The mixed usage with office and laboratory spaces and the modern sustainable energy features define a wide set of requirements for the building operator to optimize energy usage while maintaining steady occupant comfort.

2 DATA WAREHOUSE FOR ENERGY-EFFICIENT BUILDING OPERATION

Data Warehouses (DW) structure data in pre-specified *materialised views* that are defined by *dimensions* and stored in *cubes* to support data aggregation.

For example, an operator wants to analyse the energy consumption of a building and needs to know when the most energy is used (time), where it is used (location), and by which tenant (organization). This use case specifies the *dimensions* of the data warehouse respectively Time, Location, and Organization. These dimensions are used to structure and access the data in queries, for example: Give me the aggregated energy consumption of "last year" (time) for the tenant "IRUSE" (organization) in the "ERI" (location). Such aggregation queries are predefined in cubes that are spanned by dimensions and the results are pre-computed in the data warehouse, thus allowing very fast access to such results. The multi-dimensional data analysis concept and DW techniques for building performance are further detailed in Ahmed et al. (2009).

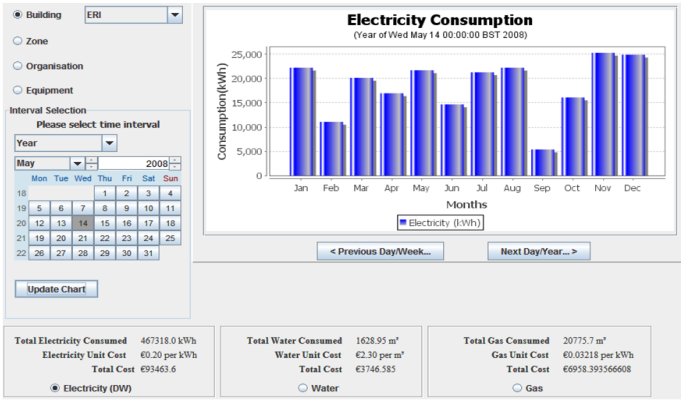


Figure 1. GUI for the building operator

Figure 1 shows the GUI implemented for the ERI DW. The three energy consumption data categories that affect the operational costs are electricity (main power board meter), natural gas (boiler and laboratory meters) and water (mains water meter). They are selectable at the bottom of the GUI. This will be extended to support the ERI´s sustainable energy systems to allow a comparison of the energy intake.

The operator uses the dimension categories to specify the data shown in the graph on the top right. The operator can select the energy consumption for a whole building, a specific zone (rooms), a tenant organization, or equipment. The calendar allows specifying the time dimension from years, to month, to single days. These dimensions enable the operator to easily analyse the building's energy consumption from top level (several years per building), down to the most detailed level (hourly per room). Due to the pre-computed queries defined by cubes, the data warehouse quickly responds with results if the operator modifies a relevant query.

3 DATA MINING CONCEPTS AND APPLICATIONS

Knowledge Discovery in Databases (KDD) and Data Mining (DM) involves processes to extract or mine knowledge from large amounts of data (Han & Kamber 2006, p. 5), providing implicit useful knowledge (Wang & Huang 2006) to address specific business problems.

Data Mining approaches can usually be categorised into descriptive and predictive algorithms. *Descriptive algorithms* on the one hand are used for exploratory data analysis to discover individual patterns, such as associations, or clusters. *Predictive algorithms* on the other hand focus on the creation of models that allow predicting observations from input data like classifications, regression models or neural networks.

Data mining has been used extensively in the medical field to solve many problems, such as the association of genes to genetically inherited diseases (Perez-Iratxeta et al. 2002). In direct marketing, data mining is able to identify likely buyers of products, advertise and promote products (Ling & Li 1998), and for products placement in shopping centres, to identify items that are likely to be purchased together. Data mining has proved successful in reducing the cost of doing business, improving profits, and increasing service quality (Apte et al. 2002). In addition, data mining supports the construction of customers' personal profile from customer transactional data (Adornavicius & Tuzhilin 2002) by the means of knowledge discovery in databases.

In buildings and energy fields, data mining approaches, like neural networks, are used in modern building automation to identify usage scenarios (Lang et al. 2007), or to estimate the energy consumption in residential buildings (Mihalakakou et al. 2002), and tropical regions (Dong et al. 2005). Characterisation of electric energy consumers was acquired using data mining (Figueiredo et al. 2005). It was also used to analyse data collected from simulations (Morbitzer et al. 2004), or wireless sensor networks (Wu & Clements-Croom 2007).

Most of these studies focus on the energy consumption of buildings, but few evaluate occupant related aspects of building performance like the thermal comfort of occupants. One reason may be, that the thermal comfort is a complex measurement itself, depending, in the case of the *Predicted Mean Vote* (PMV), on the temperature, humidity, air velocity, occupants clothing, etc. This requires complex sensor equipment for data gathering, which is not reasonable in all rooms. Data Mining can help to

solve such limitations with its predictive algorithms as this paper demonstrates.

The objective is to analyse building performance data and room thermal comfort to evaluate heating and cooling systems efficiency. The data used in this research is the historical sensed data of the ERI. The ERI has air temperature sensors in each of its 70 rooms, but possesses additional radiant temperature, humidity, and CO_2 sensors in only four rooms. To evaluate the thermal comfort for all rooms the predictive models of data mining should be used as discussed in the next sections.

4 MINING THE BUILDING PERFORMANCE DATA

Figure 2 shows the mining process of the sensor data in the ERI building. This includes data acquisition (gathering) and preparation (data access, data sampling, and data transformation), model building and evaluation (create model, test model, evaluate and interpret model), and Knowledge deployment (model apply) (Haberstroh 2008, pp. 9-12). All logical definitions and their physical implementation presented in this paper comply with Oracle Corporation Specifications for Oracle Data Miner (ODM) 11g version 1 (Oracle 2008).

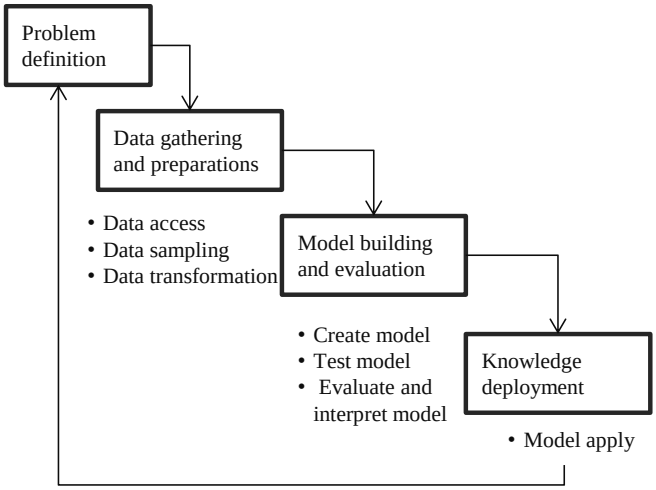


Figure 2. The process of mining the ERI sensed data stream.

4.1 *Problem definition in terms of Data Mining and Energy Management*

This section defines the problem from the energy management perspective, then converts this knowledge into a data mining problem definition and shows the preliminary plan designed to solve it.

As mentioned in Section 2, energy management is required to provide steady user comfort while reducing energy consumption. Relevant stakeholders need to evaluate HVAC system efficiency and user comfort in order to accomplish this task, while keeping the cost of this evaluation as low as possible.

We approach this problem by classifying rooms based on their thermal comfort into hot, warm, slightly warm, neutral, slightly cool, cool, and cold. The classification is based on the *Predicted Mean Vote* (PMV) as standardized in the ISO 7730 (2005). A classification model is created based on 4 rooms that have the necessary sensors available as detailed in Section 4.2.3 This model is then applied to all 70 rooms using only air temperature sensors to predict the comfort class.

4.2 Data acquisition and preparations

4.2.1 Data sources and volumes

Data processing includes cleansing, integration, and transformation of the sensed data to assure high quality (Atzmüller 2007, p. 174).

The data source for this research is a collection of storages of the ERI building performance data, as mentioned in Section 2. The ERI building is a 4500 m² "Living Laboratory" located on the campus of University Cork College, Ireland. It is equipped with multiple types of solar panels, geothermal heat pumps and an under floor heating system. Building Performance Data is provided by 180 wired sensors of the Building Management System. Additionally, a test bed for wireless sensors and actuators has been installed since April 2008 in three phases. Demonstrator 0 has been operational since June 2008. Table 1 shows the expected sensors data stream volume for the ERI building per year.

Table 1. Expected data volumes in the ERI.

Sensors	Sampling Period	Total records
180 Wired	15 minutes	6,307,200
80 Wireless	1 minutes	42,048,000
Total Volume		48,355,200

Currently, there are 190 sensors installed and working in the ERI building, with 13 different types of measurements, including indoor environment and outdoor weather conditions. These sensors are installed in 109 points in 94 rooms and spaces such as stairs way, and corridors.

4.2.2 Data collection

Data extracted and retrieved from the building's monitoring data sources is stored in a table with the attributes as listed in Table 2. These attributes are the predictors or the influences that are used to detect the room comfort class.

The data for building and testing the model used in Section 4.3 was collected for the period of 08/02/2007 to 24/04/2009 and contains 933,235 records for four rooms in the ERI building.

The data for scoring the model in Section 4.4 represents the period of 13/10/2008 to 01/02/2009 and contains 890,921 records for the air temperature and outdoor conditions for all rooms in the building.

Table 2. The predictors.

#	Attribute Name	Description
1	MEASURE_ID	A unique id to identify a sensor measure
2	ROOM_ID	A unique id to identify a room in the ERI
3	ROOM_NAME	A name to identify a room
4	ROOM_SIZE	The volume of a room
5	ROOM_FLOOR	Storey in which a room is located
6	TIME_ID	The time stamp of sensor reading
7	COMFORT_CLASS	The predicted comfort class of room
8	ROOM_TEMPERATURE	temperature measure in room
9	OUT_TEMPERATURE	Outside temperature
10	OUT_HUMIDITY	Outside humidity
11	OUT_LIGHT	Outside light
12	OUT_TOTAL_RADIATION	Outside total solar radiation
13	OUT_DIFFUSE_ RADIATION	Outside total diffuse solar radiation
14	OUT_WIND_DIRECTION	Wind direction
15	OUT_WIND_SPEED	Wind speed
16	ROOM_RAD_TEMP*	Radiant Temperature
17	ROOM_HUMIDITY*	Relative Humidity
18	ROOM_CO2*	CO_2 Concentration

*Available for 4 rooms and used only for computing the comfort class

4.2.3 Data preparations and transformation

This section shows the activity of modifying the values of some attributes and adding other values as required to present the appropriate data set for mining. There is no methodology agreed upon to prepare data for the purpose of mining, but it usually tries to identify and remove outliers, fill null-values and remove noise in the data to improve model quality.

First outliers are detected and removed. It has been found that the air temperature sensor in one room in the scoring data is broken and delivers readings between -300°C and -200°C. Second, when the Building Management System is reset it sets all measurements to zero by default. Both outliers' sources were removed from the data leaving 933,235 records for model building and 890,921 records for scoring.

However, the biggest issue concerns approximately 90% of the records per measurement (lines 8-18 in Table 2) that are NULL in the database. The reason for this is that the timestamps of the sensors are not synchronized and each sensor fills only its own column. Thus, when the air temperature sensor adds a value in the ROOM_TEMPERATURE column the other measurement columns (lines 9-18) are left empty. For data mining they need to be filled to allow the analysis of correlations.

This is done by linearly interpolating each column over the timestamp for each room. Let us assume for example the air temperature sensor in room G01 reads 20.0°C at 4:00pm and 15 minutes later 21.5°C. The relative humidity sensor adds its value at 4:05pm to the database. For this timestamp the temperature in G01 can be linearly interpolated to 20.5°C. This linear interpolation is implemented in JAVA for all continuous measurements in Table 2 for building and scoring the model.

As a last preparation step, the thermal comfort class needs to be computed for the data used for model building. The classification is based on the PMV, which is defined in the ISO 7730 and was implemented in JAVA. The PMV value is not an undisputable thermal comfort measurement (Nicol & Parsons 2002, Pfafferott et al. 2007) and other approaches try to create more general models (Yao et al. 2009). Nevertheless, the PMV was selected for this example as it shows the complexity of thermal comfort evaluation and is established. Other thermal comfort measures can be analyzed in the same way. The PMV depends on the air temperature, radiant temperature, relative humidity, air velocity, as well as occupant's clothing and activity level. Readings for the air temperature, radiant temperature, and relative humidity are available for four rooms in the database. To compute the PMV, we assume constant air velocity of 0.1m/s, which is a representative mean value for naturally ventilated offices (Moujalled 2008). At the activity level we assume office works with 1.2met. The clothing value is interpolated depending on the outside temperature between 1.0m^2K/W (indoor winter clothing at 0°C) and 0.5m^2K/W (summer clothing at 30°C).

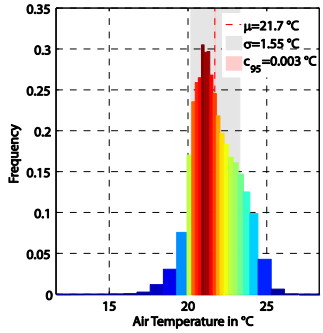
Table 3. Comfort classes based on the PMV.

Comfort Class	Classification	No. in Data	Percentage in Data
Hot	$3.5 > PMV \geq 2.5$	0	0.0%
Warm	$2.5 > PMV \geq 1.5$	4	0.0%
Slightly Warm	$1.5 > PMV \geq 0.5$	8,948	1.0%
Neutral	$0.5 > PMV \geq -0.5$	772,072	82.7%
Slightly Cool	$-0.5 > PMV \geq -1.5$	150,227	16.1%
Cool	$-1.5 > PMV \geq -2.5$	1,984	0.2%
Cold	$-2.5 > PMV \geq -3.5$	0	0.0%
OutOfRange	otherwise	0	0.0%

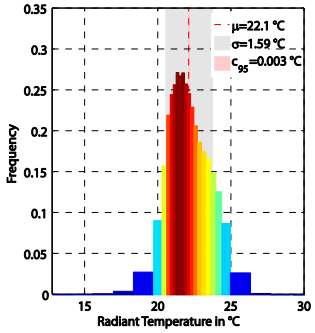
The comfort class is assigned from the PMV value according to the classification in Table 3. The table lists also the resulting numbers of entries in each class. The distribution of the PMV and the room measurements are displayed in Figure 3 for comparison. The distributions of the PMV values are about the same for all four rooms.

Figure 4 shows the results of the attribute importance analysis of the Oracle Data Miner run on the computed comfort classes for the model building

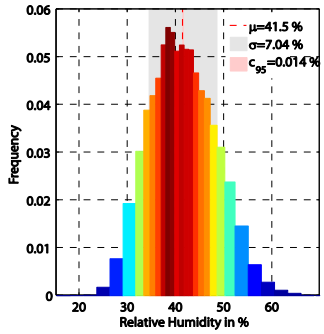
data. *Attribute Importance* identifies the subset of attributes relevant for classification using a Minimum Description Length Algorithm (Oracle 2008). It is obvious that the PMV and the related Percentage of Persons Dissatisfied (PPD) have the biggest influence on the comfort class. The air and radiant temperatures are next in rank of importance. Other values are less important for the comfort classification.



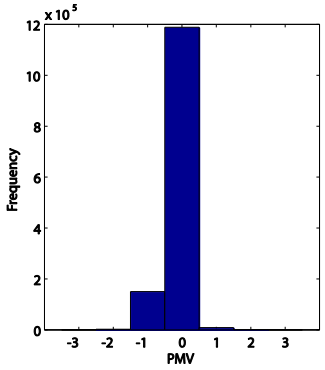
a) Room air temperature



b) Room radiant temperature



c) Room relative humidity.



d) Room PMV.

Figure 3. Histograms of various measures from 4 rooms. (μ - mean value; σ - standard deviation; c_{95} – 95% confidence interval)

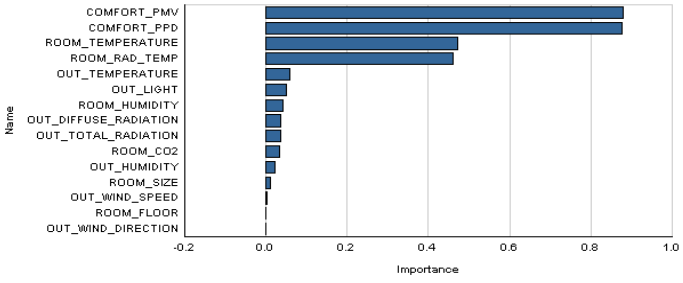


Figure 4. Influences of the indoor measures in room comfort.

This is relevant for the model building in the next step, as the PMV, PPD, radiance temperature, relative humidity, and CO_2 are removed, as they are not available in the other rooms on which the model should be applied to. We assume that this is feasible, as the room radiance temperature is strongly correlated to the room air temperature (compare Figure 3a and 3b, the room humidity is correlated to the outside humidity, and the clothing level was related to the outside temperature during the PMV computation. Several tests in the next section will show if this assumption is correct.

4.3 Building and evaluating the comfort model

Building a data mining model is the process of finding the best algorithm or technique, by which the building sensed data is analysed and represented as patterns and rules (Harinath & Quinn 2006, p. 485).

The following shows how to classify room comfort. This is an overview of building, testing, and scoring a classification model.

Classification is a model or a classifier that is constructed to predict the categorical label of a room in a building (Han & Kamber 2006, p. 286). These classes are defined in Section 4.2.3. Classification mining function uses different algorithms such as decision trees, Naïve Bayes, and support vector machines.

As the attributes in Table 2 are unconditional this makes Naïve Bayes the optimal algorithm (Fielding 2007, p. 99) to detect room comfort in buildings in this case. Naïve Bayes is a probabilistic classifier that uses Bayesian theory. It simplifies the learning by assuming that the attributes in Table 2 are independent (Abellan et al. 2007) given the room comfort class as the variable to classify. Decision trees and support vector machines resulted in poor models.

In the setting phase to build the model, the cool label has been used as the preferred target value. Data split into two subsets of 60% and 40% for training and testing the models. The 40% is called a holdout sample or a test dataset. The sampling process was disabled, as the model building time was acceptable for our data size. The model was tuned towards a maximum average accuracy that creates a model that is good in predicting all labels (Huang et al. 2008).

During the building process the model learns from the sensed data how to distinguish between comfort classes in order to predict the same classes when the model is applied to other rooms. The test metrics of ODM, which are detailed in the following sections, allow evaluation of the model's quality (Maimon & Rokach 2005, p. 1241).

4.3.1 Predictive confidence

Predictive confidence is a visual indication of the effectiveness of this model compared to a random guess of the rooms' comfort class. It is a validation of the ability of the model to generalize what it learned in a different data set (Fernández 2003, p. 152). If the needle in Figure 5 points to the lowest point on left of the dial, then the model is no better than a random guess (Haberstroh 2008, p. 85). The comfort detection model developed in this study shows 85.28% predictive improvement over a random guess in predicting rooms' comfort class. In

comparison, a classification model taking also the rooms' humidity, CO2 and radiant temperature into account reaches a predictive confidence of 89.44%. If only the rooms' air temperature is used for classification, the predictive confidence reduces to 74.75%. This shows on the one hand the high importance of the rooms' air temperature for the comfort class. On the other hand, this demonstrates also that the other values considered in this study, like outside measurements and room size, improve the model significantly.

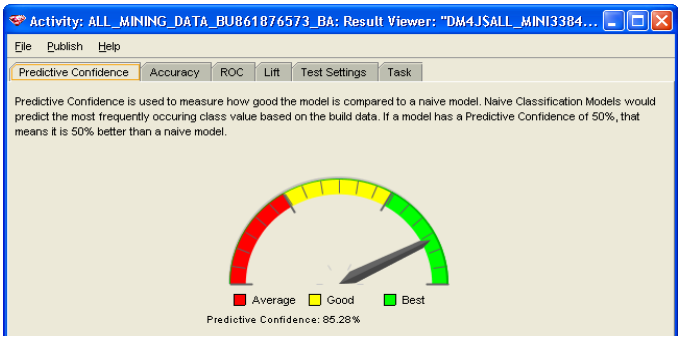


Figure 5. The predictive confidence of the model.

4.3.2 Model accuracy

Model accuracy shows the several interpretations of the fault detecting model ability in predicting the class when applied to the test data.

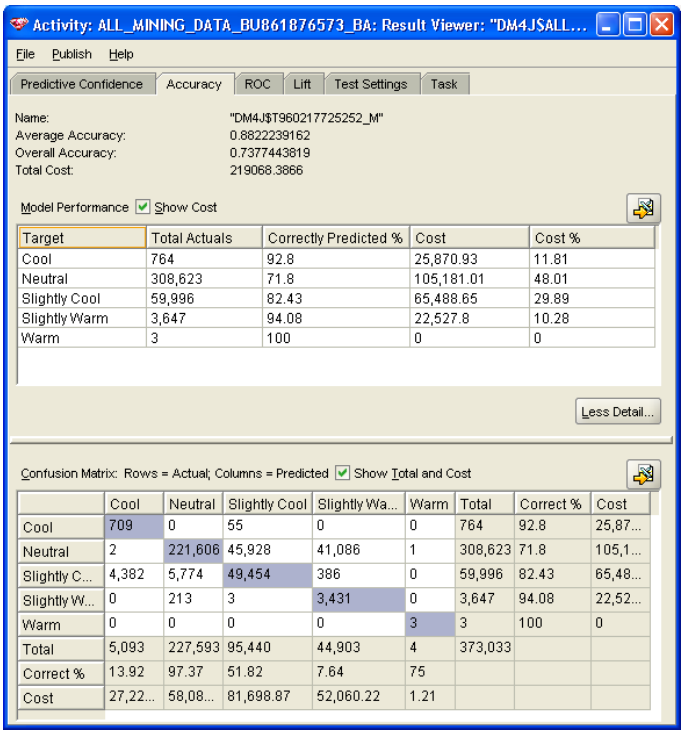


Figure 6. Model accuracy and the confusion matrix.

Figure 6 shows the model accuracy for the comfort classification. The table on the top shows the percentage of values correctly predicted per class. For example, there are 308,623 cases with a comfort class 'neutral' and the model predicts 71.8% of them correct. The cost is an indication of damage done by incorrect prediction (Berry & Linoff 2004, p. 79), and it is a valuable metric for model comparisons. The displayed model was the best model we could develop, with the lowest cost of predicting rooms' comfort classes.

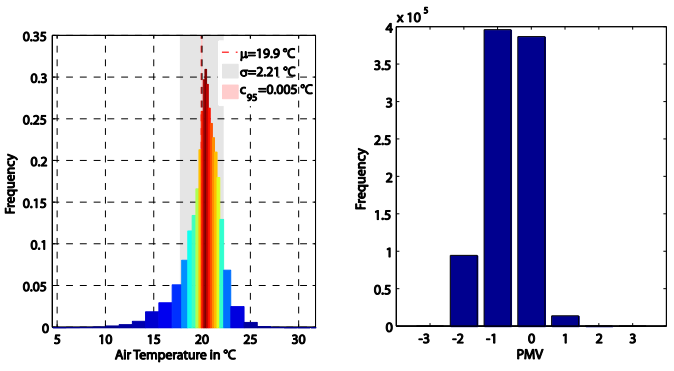
The type of errors expected from this model is shown on the *confusion matrix* in the lower table in Figure 6. Actual (correct) values of the classes are represented by rows and compared against the predictions made by the model in columns. The numbers tell how many classes were correctly predicted or misinterpreted as another class. For example, the first row in Figure 6 indicates that, of the samples with the actual comfort class 'cool', 709 cases were correctly predicted and 55 cases were predicted incorrectly as 'slightly cool'.

To interpret the confusion matrix, incorrect prediction variations are usually placed next to the correct classes, i.e. the 'neutral' class is either predicted incorrectly as 'slightly cool' or 'slightly warm'. The rare classes 'warm' and 'cool' have a high percentage of correct prediction. The reason is probably that they are characterized by extreme air temperatures. However, the low number of samples do not allow generalisation in so far as the classes will also be detected correctly in other data. As the building data contained no 'hot' and 'cold' cases the model will not be able to classify these classes.

4.4 Knowledge deployment

The created model can be applied to any building performance data that has the same structure and format, to predict the comfort class. The applying activity is sometimes referred to as scoring the model (Giovinazzo 2002, p. 168) that uses the model in a different data set to predict the classification.

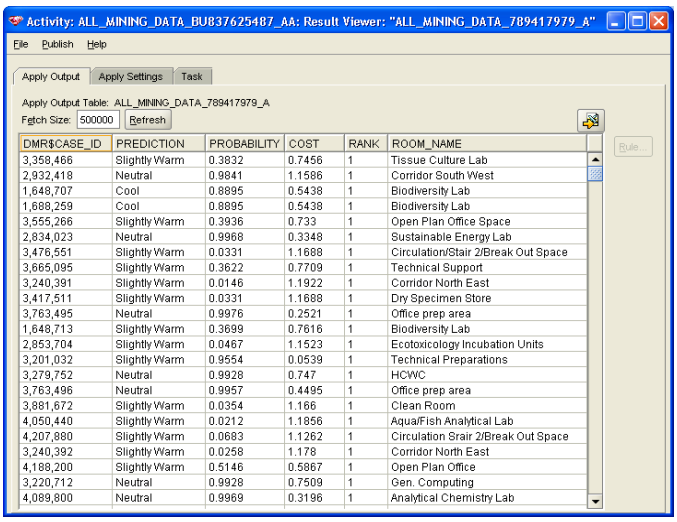
This is done for all 70 rooms excluding the room with the broken temperature sensor, which was cleaned out as explained in Section 4.2.3. The new model allows predicting the thermal comfort class based only on the rooms' air temperature and the buildings outside conditions.



a) Room air temperature. b) Room PMV.

Figure 7. Histograms of measures from all rooms.

The distributions for the air temperature of these rooms and the predicted PMV values are shown in Figure 7. The mean air temperature (μ) is 19.9°C and slightly lower than the 21.7°C for the four rooms' data used for model (see Figure 3a). The standard deviation increases from 1.6°C to 2.2°C as the added rooms increase the variance. This results in a broader PMV distribution in Figure 7b in comparison to Figure 3c, with significantly more 'slightly cool' and 'cool' values.



DMR$CASE_ID	PREDICTION	PROBABILITY	COST	RANK	ROOM_NAME
3,358,466	Slightly Warm	0.3832	0.7456	1	Tissue Culture Lab
2,932,418	Neutral	0.9841	1.1586	1	Corridor South West
1,648,707	Cool	0.8895	0.5438	1	Biodiversity Lab
1,688,259	Cool	0.8895	0.5438	1	Biodiversity Lab
3,555,266	Slightly Warm	0.3936	0.733	1	Open Plan Office Space
2,834,023	Neutral	0.9968	0.3348	1	Sustainable Energy Lab
3,476,551	Slightly Warm	0.0331	1.1688	1	Circulation/Stair 2/Break Out Space
3,665,095	Slightly Warm	0.3622	0.7709	1	Technical Support
3,240,391	Slightly Warm	0.0146	1.1922	1	Corridor North East
3,417,511	Slightly Warm	0.0331	1.1688	1	Dry Specimen Store
3,763,495	Neutral	0.9976	0.2521	1	Office prep area
1,648,713	Slightly Warm	0.3699	0.7616	1	Biodiversity Lab
2,853,704	Slightly Warm	0.0467	1.1523	1	Ecotoxicology Incubation Units
3,201,032	Slightly Warm	0.9554	0.0539	1	Technical Preparations
3,279,752	Neutral	0.9928	0.747	1	HCWC
3,763,496	Neutral	0.9957	0.4495	1	Office prep area
3,881,672	Slightly Warm	0.0354	1.166	1	Clean Room
4,050,440	Slightly Warm	0.0212	1.1856	1	Aqua/Fish Analytical Lab
4,207,880	Slightly Warm	0.0683	1.1262	1	Circulation Srair 2/Break Out Space
3,240,392	Slightly Warm	0.0258	1.178	1	Corridor North East
4,188,200	Slightly Warm	0.5146	0.5867	1	Open Plan Office
3,220,712	Neutral	0.9928	0.7509	1	Gen. Computing
4,089,800	Neutral	0.9969	0.3196	1	Analytical Chemistry Lab

Figure 8. Sample of output table for applying the model.

A sample of the output table of applying the model is displayed in Figure 8. The sample table shows each row with the identifier, prediction of the most likely class, the probability that this is the right guess; the cost of incorrect prediction, and the rank to categorize predictions. The room name was added to ease readability.

The model estimates to make correct predictions with mean probabilities of 78%. The 'neutral' label is usually predicted with 97% mean probability, 'slightly cool' with 76%, 'slightly warm' with 21%, and 'cool' and 'warm' with 15% mean probability. The reason for this distribution is that the data used for building the model contained mostly cases for 'neutral', which increases the model quality for this case, but the lack of data for the other cases reduces their model quality.

4.5 Knowledge gained and interpretation

As a last step, the PMV distribution for a room was analysed to identify the rooms with an emphasis on not 'neutral' comfort level. See Figure 9 for a location of the rooms. 40 rooms out of 70 were identified as having mainly 'slightly cool' comfort level, and 5 rooms had a 'cool' comfort level for more than 30% of the cases. Four of these five rooms are located at the south facade on ground level and three have exterior doors. The ground floor has the highest number of rooms with 'neutral' comfort. One room in the middle of the floor shows abnormal behaviour that should be investigated, as the room has more than 30% 'cool' comfort level in contrast to its neighbour rooms.

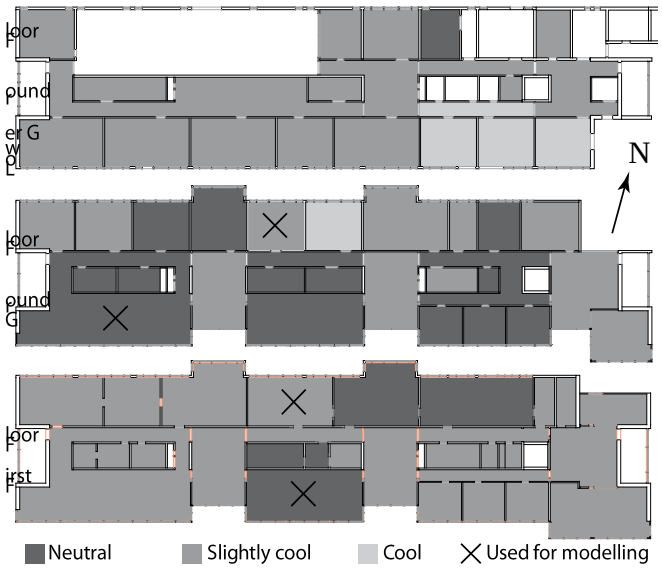


Figure 9. Comfort levels of the rooms.

In general, the thermal comfort for the scored winter period was 'slightly cool'. The set point temperature for all rooms was 20°C, which represents the mean temperature value as shown in Figure 7a. However, to provide a better thermal comfort, the set point should be higher.

Office hours were not considered during the analysis. The mean air temperature varies in the scoring data by 1.5°C reaching the minimum of 19.0°C at 2am and the maximum of 20.5°C at 4pm.

5 CONCLUSION

Two approaches were introduced to analyse building performance data for energy-efficient buildings.

The data warehouse solution provides a single repository for building performance data, creates sophisticated energy aggregations, and provides friendly user interfaces.

The data mining model automates and eases evaluating building thermal comfort, while reducing the cost of monitoring equipment. The process from data acquisition, preparation, model building, to knowledge deployment was examined using real data from the ERI. The results show that the approach is feasible, but more data is needed to train the model for less frequent classes like 'hot' and 'cold'.

Implementing data mining techniques to building sensed data will help in stabilising rooms' preferences while optimising energy usage. Therefore, the correlations between the building energy usage and thermal comfort will be further examined with a

special focus on the sustainable energy sources of the ERI. Another future research topic will be the development of mining models for fault detection and diagnosis as well as mining models that consider human comfort feedback along with other influences in room states, such as the structural properties of the building and its geometrical specifications. The extensions of the ERI with a further 80 wireless sensors will increase the data set for analysis and will also provide more validation data for this model. These solutions are used by the ITOBO (2007) project to increase the value of energy-efficient smart buildings.

6 ACKNOWLEDGEMENT

Work in the Strategic Research Cluster 'ITOBO' is funded by Science Foundation Ireland and additional contributions from 5 industry partners. Joern Ploennigs is as Feodor Lynen Fellow in Cork and wants to thank the Humboldt-Foundation and the German BMBF for their support.

The authors thank Paul Stack, Luke Allan, Brian Cahill, Civil Engineering UCC; Anika Schumann, Cork Constraint Computation Centre; and Haithum Elhadi, U.S. Telecom and Illinois Institute of Technology for their contribution to this research.

7 REFERENCES

Abellan, J., Cano, A., Masegosa, A. R., & Moral, S. 2007. A Semi-Naive Bayes Classifier with Grouping of Cases. In K. Mellouli (Ed.), *9th European Conference, ECSQARU* (pp. 477-488). Hammamet, Tunisia: Springer.

Adornavicius, G., & Tuzhilin, A. 2002. Using data mining methods to build customer profiles. *IEEE Computer* 34 (2): 74-82.

Ahmed, A., Menzel, K., Ploennigs, J., & Cahill, B. 2009. Aspects of Multi-dimensional Data Analysis of Building Performance Data Management. *16th European Group for Intelligent Computing in Engineering International Workshop.* Berlin, Germany, accepted.

Apte, C., Liu, B., Pednault, E. P., & Smyth, P. 2002. Business Application of Data Mining. *Communications of the ACM* 45 (8): 49-53.

Atzmüller, M. 2007. *Knowledge-intensive Subgroup Mining: Techniques for Automatic and Interactive Discovery.* IOS Press.

Augenbroe, G., Park, C. S. 2005. Quantification methods of technical building performance. *Building Research and Information* 33 (2): 159-72.

Berry, M. J., & Linoff, G. 2004. *Data mining techniques: for marketing, sales, and customer relationship management.* John Wiley and Sons.

Capehar, B. L., Turner, W. C., & Kennedy, W. J. 2008. *Guide to Energy Management.* The Fairmont Press.

Crawley, D. B., Hand, J. W., Kummert, M., & Griffith, B. T. 2008. Contrasting the capabilities of building energy performance simulation programs. *Building and Environment* 43 (4): 661-673 .

Dong, B., Cao, C., & Lee, S. E. 2005. Applying support vector machines to predict building energy consumption in tropical region. *Energy and Buildings* 37 (5): 545-553.

ERI 2002. Environmental Research Institute. Cork, Ireland: University College Cork, http://eri.ucc.ie.

Fernández, G. 2003. *Data mining using SAS applications.* CRC Press.

Fielding, A. 2007. *Cluster and classification techniques for the biosciences.* Cambridge University Press.

Figueiredo, V., Rodrigues, F., Vale, Z., & Gouveia, J. B. 2005. *IEEE Transaction on Power Systems* 20 (2): 596-602.

Giovinazzo, W. A. 2002. *Internet-enabled business intelligence.* Prentice Hall PTR.

Haberstroh, R. 2008. *Oracle Data Mining Tutorial for Oracle Data Mining 11g Release 1.* Oracle.

Han, J., & Kamber, M. 2006. *Data mining: concepts and techniques* (2 ed.). Morgan Kaufmann.

Harinath, S., & Quinn, S. R. 2006. *Professional SQL server analysis services 2005 with MDX.* John Wiley and Sons.

Huang, B., Cai, Z., Gu, Q., & Chen, C. 2008. Using Support Vector Regression for Classification. *4th International Conference on Advanced Data Mining and Applications* (pp. 581-588). Chengdu, China: Springer.

ISO 7730:2005. Ergonomics of the thermal environment - Analytical determination and interpretation of thermal comfort using calculation of the PMV and PPD indices and local thermal comfort criteria.

ITOBO 2007. Information & Communication Technology for Sustainable and Optimised Building Operation. Cork, Ireland: http://zuse.ucc.ie/itobo/.

Lane, P. 2007. *Data Warehousing Guide, 11g Release 1 (11.1),* Oracle Data Base, Oracle.

Lang, R., Bruckner, D., Pratl, G., Velik, R., & Deutsch, T. 2007. Scenario recognition in modern building automation. *7th IFAC International Conference on Fieldbuses & Networks in Industrial & Embedded Systems,* (pp. 305-312).

Ling, C. X., & Li, C. 1998. Data Mining for Direct Marketing: Problems and Solutions. *4th International Conference on Knowledge Discovery and Data Mining*, (pp. 73-79).

Maimon, O. Z., & Rokach, L. 2005. *Data mining and knowledge discovery handbook.* Springer Science & Business.

McCue, C. 2006. *Data mining and predictive analysis: intelligence gathering and crime analysis.* Butterworth-Heinemann.

Menzel, K., Pesch, D., O'Flynn, B., Keane, M., & O'Mathuna, C. 2008. Towards a Wireless Sensor Platform for Energy Efficient Building Operation. *12th International conference on Computing in Civil and Building Engineering* (pp. 381-386). Beijing, China : Elsevier B.V.

Metz, B. 2007. *IPCC Fourth Assessment Report on the mitigation of climate change for researchers, students, and policymakers.* University Press.

Mihalakakou, G., Santamouris, M., & Tsangrassoulis, A. 2002. On the energy consumption in residential buildings. *Energy and Buildings* 34 (7): 727-736.

Morbitzer, C. and Strachan, P. and Simpson, C. 2004. Data mining analysis of building simulation performance data. *Building Services Engineering Research and Technology* 35 (3): 253–267.

Moujalled, B., Cantin, R., & Guarracin, G. 2008. Comparison of thermal comfort algorithms in naturally ventilated office buildings. *Energy and Buildings* 40 (12): 2215–2223.

Nicol, F., Parsons, K. 2002, Special issue on thermal comfort standards, *Energy and Buildings* 34 (6): 529-685.

Oracle. 2008. *Oracle Data Mining Concepts.* Oracle.

Perez-Iratxeta, C., Bork, P., & Andrade, M. A. 2002. Association of genes to genetically inherited diseases using data mining. *Nature Genetics* 31: 316-319.

Pfafferott, J. U., Herkel, S., Kalz, D. E., Zeuschner, A. 2007 Comparison of low-energy office buildings in summer using different thermal comfort criteria, *Energy and Buildings* 39 (7): 750-757.

Rob, P., Coronely, C., & Crockett, K. 2008. *Data Bases Systems: Design, Implementation and Management*. Cengage Learning EMEA.

Stackowiak, R., Rayman, J., & Greenwald, R. 2007. *Oracle data warehousing and business intelligence solutions.* John Wiley and Sons.

Wang, X., & Huang, J. Z. 2006. A Cased-Based Data Mining Platform. In G. J. Williams, & S. J. Simoff, *A State of the Art Survey, Data mining: theory, methodology, techniques, and applications* (pp. 28-39). Springer Science & Business.

Witten, I. H., & Frank, E. 2005. *Data mining: practical machine learning tools and techniques* (2 ed.). Morgan Kaufmann.

Wu, S., Clements-Croom, D. 2007 Understanding the indoor environment through mining sensory data—A case study. *Energy and Buildings* 39 (11): 1183–1191.

Yao, R., Li, B., Liu, J. 2009. A theoretical adaptive model of thermal comfort - Adaptive Predicted Mean Vote (aPMV), *Building and Environment* 44 (10): 2089-2096.