ISEIS

Journal of Environmental Informatics 13(1) 45-55 (2009)

Journal of Environmental Informatics

www.iseis.org/jei

Predictive System for Monitoring Regional Visitor Attendance Levels in Large Recreational Areas

T. Räsänen*, H. Niska, T. Hiltunen, J. Tiirikainen and M. Kolehmainen

Research Group of Environmental Informatics, Department of Environmental Sciences, University of Kuopio, P.O. Box 1627, FIN-70211, Kuopio, Finland

Received 18 March 2008; revised 12 November 2008; accepted 18 November 2008; published online 12 March 2009

ABSTRACT. The number of people in a certain place in desired time period is one of the main questions in many monitoring and management applications. Such information is needed also in tourism which has been one of most growing business areas in recent years and it have become a more important area of the service sector. Thus effective and sustainable management is needed. Regional visitor monitoring provides general tools for managers and decision-makers to handle multidimensional growth and the development of tourism business. Furthermore, predictive information is a key for solving many environmentally related daily problems like minimizing raw material loss. In this study, we present continuous approach for monitoring and predicting regional visitor attendance levels using inferred data and self-organizing map. The data used in study contained variables describing regional mobile telecommunication events, weather conditions, traffic density, restaurant sales and the use of accommodation facilities. The self-organizing map (SOM) was used to integrate the variables into a combined regional attendance index and multilayer perceptron was applied to predict the number of visitors. The proposed method was tested and an online modelling system created using real data gathered from the recreational area of Tahko. In general, the results showed that the method is suitable for describing a real regional situation and seasonal variations in visitor attendance levels. Moreover, the results indicated that mobile telecommunication data improve the modelling of daily visitors at a more detailed level. Presented online modelling system were applied, during the research project, to optimize number of operating taxi vehicles, plan opening hours and times of grocery shop and reschedule the collection of municipal wastes.

Keywords: positioning of masses of people, mobile telecommunications data, regional modeling, self-organizing map, tourism, visitor monitoring

1. Introduction

The number of people in a certain place in desired time period is one of the main questions in many monitoring and management applications. Nevertheless, answer of this question is not obvious in cases where people are widely spread inside large areas. Tourism has become an important area of the service sector, providing new opportunities for regional economic development and increasing the proportion of the discretionary income of tourists spent in this sector (Lim and McAleer, 2005). Moreover, the popularity of outdoor sports, leisure activities and nature tourism has been growing in recent years (Krämer and Roth, 2002) and in many places natural forests and wildlife have been placed with recreational areas that offer numerous experiences of nature and leisure activities. This increasing use of the landscape has aroused severe problems in some affected areas, including local conflicts between commercial activities and the use of nature (Krämer and Roth, 2002). Sustainable and systematic management is therefore needed to guide activities in recreational areas.

The monitoring of vegetation and wildlife in recreational and protected areas has been widely discussed and researched. A scientific interest in observing the development of regional ecosystems has often been the driving force for creating monitoring systems. On the other hand, systematic monitoring of recreational uses, the effects of leisure activities and visitor flows is rarely carried out (Muhar et al., 2002). A comprehensive understanding of their use is very important for the sustainable and effective management of recreational areas, and detailed regional information on leisure-time and recreational usage can make it possible to combine knowledge from the fields of natural science and sociology in order to ensure the ecologically and economically sustainable management of recreational areas (Heywood, 1993; Hornback and Eagles, 1999).

On the other hand, information on visitor numbers and visitor attendance levels over time is also important for a variety of strategic and operational planning tasks in tourist centres, conservation areas and natural parks. Accurate information is also needed to assess business plans and actions relative to sea-

* Corresponding author. Tel.: +358 17 162337; fax: +358 17 163191.
E-mail address: teemu.rasanen@uku.fi (T. Räsänen).

ISSN: 1726-2135 print/1684-8799 online
 doi:10.3808/jei.200900139

sonal numbers of visitors. Systematic visitor monitoring can provide fundamental visitor management data which can be used in many important applications such as visitor flow modelling (Cottrell, 2002; Cessford et al., 2002; Eagles et al., 2001).

Moreover, it is also possible by using accurate information to detect regional changes over time and determine whether these changes are due to natural causes or to stress caused by human activities (Parks Canada, 1994). When unusual events occur, accurate and comparable historical data are needed to deal with the unexpected consequences. In other words, current visitation data produced by systematic, continuous monitoring systems may be of critical importance at some later point in time (Hornback and Eagles, 1999).

A system of basic visitor monitoring or public use reporting should contain methods for: (1) data gathering, (2) summarization of data, (3) analysis of data, and (4) interpretation of data for management action (Hornback and Eagles, 1999). Visitor monitoring systems should provide wide-ranging, up-to-date seasonal information about visitor numbers, attendance levels and trends. Recreational area managers and commercial actors, for example, need estimates of visitor attendance levels for (1) the preparation of staffing plans (Brandenburg and Ploner, 2002), (2) planning of the usage of raw materials and groceries to minimize loss and waste, (3) finding the optimal timing (best seasons) for marketing and certain services, and (4) making sustainable decisions about nature use and protection.

The collection of visitor count data in conservation or recreational areas is difficult, and visitor monitoring can be an opportunistic exercise involving a combination of counting methods and techniques. Traditional techniques can be classified into three types: (1) direct observations using staff observers or camera recordings at sites, (2) on-site counters or other devices to record visitor presence, and (3) inferred counts, i.e. the use of other data to obtain on-site estimates. Combinations of these approaches are often used. Direct observations and on-site counters are in general accurate if the placement and location of the counter is relevant with respect to the movement of visitors. These methods are always tied to a particular location and often need specific on-site structures and electricity. All these methods have advantages and disadvantages, and the selection of one particular visitor counting approach and method will always be a compromise between accuracy and practical measurement capacity (Cessford et al., 2002).

However, the main aim was to develop predictive and continuous approach for monitoring visitor attendance levels using inferred data sources and computationally intelligent methods, like Self-organizing Map (SOM) and Multilayer Perceptron (MLP). A further aim was to point out novel ways of using mobile telecommunications data for environmental applications. The modelling and computations were done using the Matlab environment (The Mathworks Inc.), especially SOM Toolbox (Helsinki University of Technology). The research and methodological development work was carried out in accordance with the ideology of environmental informatics by applying information technology to environmental issues using data-driven methods (Kolehmainen, 2004; Green and Klomp, 1998; Page and Rautenstrauch, 2001).

2. Materials and Methods

All management is dependent upon information and the better the quality of information; the better the opportunity for good management. The more detailed information about the visitors and their activities enables managers to deal with the challenge of increasing volumes of tourism (Hornback and Eagles, 1999) and the problems of daily operations. The continuous information of the current visitor attendance and near future forecast are useful tool for handling many problems due to widely varying visitor amount like raw material loss in restaurants or grocery shops, suddenly increased waste amounts and over- or undersized personnel. Furthermore, collected and easily available historic information gives a great opportunity to evaluate effects of different kind of development actions. The area of Tahko is typical growing recreational area where above problems occur. Therefore area was suitable for testing and evaluating performance of proposed system and it was carried out in close collaboration with regional entrepreneurs, municipal authorities and other actors.

2.1. The Tahko Area

The Tahko area, containing a nature and conservation and recreational areas, is situated in Eastern Finland, not far from the town of Nilsiä. It is about 60 km north-east of the city of Kuopio and 450 km north-east of the capital, Helsinki (Figure 1). The area belongs to the westernmost part of the Siberian taiga, where the summers are fairly warm and the water in the lakes warms up quickly to swimming temperatures. It is a hilly area with a maximum altitude of 314 metres above sea level, and has a lush flora and abundant stretches of water. July is the warmest month of the year, with an average temperature of +21°C, and January is usually the coldest month, with an average temperature of −15°C. The area contains three Natura 2000 sites which are dominated by ecologically valuable Boreal herb-rich forests, bogs and aquatic environments.

The recreational area of Tahko is well-known as a holiday centre. It is the biggest combined downhill and cross-country ski resort in the southern half of Finland, and has facilities for many other outdoor activities such as snowmobiling, snowshoeing, hiking, boating, canoeing, golfing, rock climbing etc., together with accommodation, restaurants and leisure services in winter and summer. Its accommodation capacity is approximately 5000 beds (Tahkovuori Ltd., 2005) and it receives some 169,000 visitors per year according to downhill skiing ticket sales (Santasalo, 2007).

2.2. The Data Used in Study

The method proposed for modelling regional visitor attendance is based on time-series data gathered from the following sources in the Tahko area: (1) mobile telecommunications event data from the mobile network (GSM) of Finnet Ltd., (2) accommodation data from Tahkovuori Ltd., (3) tra-

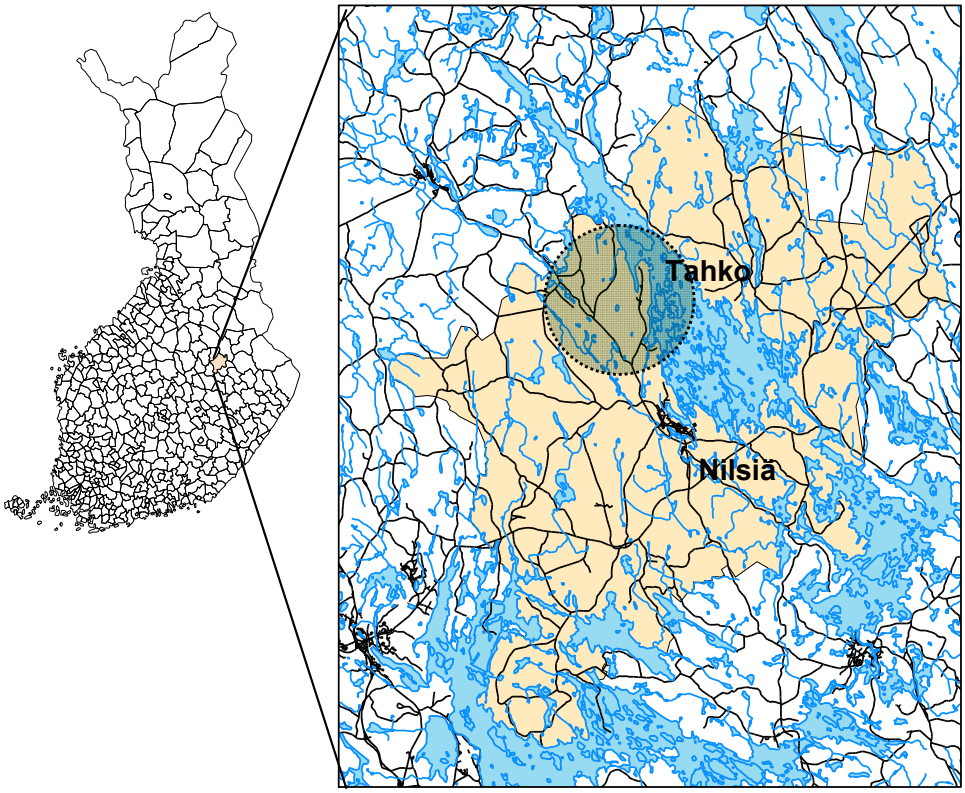


Figure 1. The Tahko area (located near the small town of Nilsiä in Eastern Finland and surrounded by lakes).

ffic density data from the National Road Administration's measuring point, (4) daily restaurant sales data from TahkoChalet Ltd., and (5) weather data from the Finnish Meteorological Institute.

The penetration of mobile phones in Finland 2004 was 4,999,060 which means that 95 citizens of 100 has such phone (Statistics Finland, 2007). From this point of view, we can assume that most of the tourists have also mobile phone with them. Mobile telecommunication data were collected from seven base transceiver stations (BTS) which are part of the mobile (GSM) network of Finnet Ltd., which has an approximately 20% market share in the Finnish mobile phone sector. The area covered by each base transceiver station is called a cell (Drane et al., 1998) and these seven cells cover most of the Tahko area.

The telecommunications data is normally used for customer billing and management purposes. Furthermore, the cell-based positioning is typically used to determine the geographical location of individual mobile users, but in this case the data and cell-based information was used in a novel way to estimate the number of people in a certain region at a given time. Simplified demonstration of positioning of masses of people used in this study is illustrated in Figure 2.

Recently, mobile phone data has been used also to understand individual human mobility patterns. The location information, based on recording a call or text messages initiation place, for 100,000 individuals was collected for a six-month period and the typical travel patterns were solved. Human mobility could impact for example phenomena's like; epidemic prevention to emergency response, urban planning and agent based modelling (González et al., 2008).

The raw data contained date, time, area code, cell identification code and type for every telecommunication event, i.e. all outgoing and incoming calls, general packet radio service data transfers (GPRS) and short message services (SMS). The raw data was transformed into four different hourly time-series by summing the events in a particular hour. It is well-known that mobile data based on cell-based positioning can be used to determine the geographical location of individual mobile users, but in this case the cell-based information was used in a novel way to determine the volume of mobile users, i.e. the number of people in certain region at given time.

The accommodation data came from the online booking systems maintained by Tahkovuori Ltd., which serves as the central booking agency for the area and manages approximately 90% of all information on its accommodation bookings. The raw data contained daily information about usage and reservations (in terms of numbers of beds, distinguishing separately the numbers of adults and children).

Traffic density data was collected by the Finnish Road Administration from detectors located beside the nearest main road and transferred to the modelling server via an internet-

based online system. The Finnish Road Administration has a total of 300 such measuring points all over Finland. In this case the detector was located beside the larger of the two main roads entering the Tahko area. In principle it detects all vehicles that pass by, but in this case only the hourly numbers of light vehicles were used for modelling purposes. The raw data contained the number of vehicles per hour in both directions.

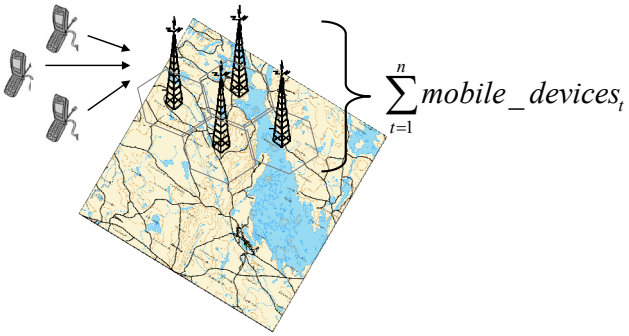


Figure 2. Simplified demonstration of using mobile telecommunication network to locate masses of people in a certain region (the number of mobile devices in each network cell is used to estimate number of people).

Restaurant sales data was gathered from the daily sales reports of one of the main restaurants in the ski resort, Restaurant Piazza, which is owned by TahkoChalet Ltd. and is located in a central position in the resort. The raw data contained daily sales in euros. Weather data was collected from the nearest weather observation station maintained by the Finnish Meteorological Institute, at Rissala, approximately 25 kilometres south of the Tahko area. The raw data was recorded at 3-hourly intervals and contained the following variables: date, time, temperature, humidity, wind direction, wind speed, cloudiness and current weather code. In this case only the temperature and humidity variables were used. The raw data were interpolated to provide one-hour resolution before modelling.

Finally, after pre-processing of the data, we had data set containing 17 continuous time-series (with a resolution of one hour) for the one-year period 1st Dec. 2003 to 30th Nov. 2004. These variables are presented in Table 1 and linear correlations between them are illustrated in Table2. Correlation between binary variables, marked with (*) in Table 2, and nominal variables was calculated using point-biserial correlation coefficient.

2.3. Data Processing Chain

The presented method for the modelling of regional visitor attendance levels employs inferred data on the tourism activities and combination of several data processing and computationally intelligent methods. These computational methods are typically used in data mining (Hand, 2001). The automated data processing chain is summarized in Figure 3.

The data processing chain consists of seven phases, starting with data collection and pre-processing. The data was collected via the internet, by email or using special Java Web Services interfaces, depending on the source data management system. The pre-processing step included the replacement of missing values, the creation of derivative variables and transformations, e.g. variance scaling of variables.

Table 1. Variables Used in the Model

Time variables:	
yd_sin	Day of year (1-365) (transformed to continuous form using the sine-function)
yd_cos	Day of year (1-365) (transformed to continuous form using the cosine function)
holnum_1	Binary variable for a normal workday
holnum_2	Binary variable for a Saturday or eve of a feast-day
holnum_3	Binary variable for a Sunday or feast-day
sin_hour	Sine of hour
cos_hour	Cosine of hour
Weather variables:	
temp	Temperature
hum	Humidity
Cell-based mobile telecommunication variables:	
calls	Number of outgoing calls
rcalls	Number of calls received
grps	Number of grps data transfer events
sms	Number of sms messages
Accommodation variables:	
adults	Daily number of adults with accommodation bookings
childs	Daily number of children with accommodation bookings
Restaurant sales variables:	
restaurant	Daily sale in euros
Traffic density variables:	
traffic_sum	Number of light vehicles

2.4. Regional Model

The regional model for visitor attendance level was constructed employing the SOM technique, one of the best-known unsupervised neural learning algorithms, which produces a continuous two-dimensional mapping from a multidimensional input space. The input vectors, which have common features, are projected to the same area of the map, which is consisted of neurons. Each neuron is associated with an n-dimensional reference vector, which provides a link between the output and input spaces. During learning, the input data vector is mapped onto a particular neuron based on the minimal n-dimensional distance between the input vector and the reference vectors of the neurons. Then the reference vectors of the activated neurons are updated. When the trained map is applied, the best matching units (BMU's) are calculated using these reference vectors. In this unsupervised methodology, input data can be used to construct the SOM without previous a priori knowledge. For a more detailed review of the SOM technique, the reader is referred to Kohonen (1997). Applications of SOM in the

Table 2. Linear Correlations between Variables

	ydsin	ydcos	holnum1*	holnum2*	holnum3*	sinhour	coshour	temp
ydsin	1.00	0.02	0.08	-0.04	-0.06	0.00	0.00	0.82
ydcos	0.02	1.00	0.00	0.00	0.00	0.00	0.00	-0.29
holnum1*	0.08	0.00	1.00	-0.63	-0.65	0.00	0.00	0.03
holnum2*	-0.04	0.00	-0.63	1.00	-0.18	0.00	0.00	0.00
holnum3*	-0.06	0.00	-0.65	-0.18	1.00	0.00	0.00	-0.04
sinhour	0.00	0.00	0.00	0.00	0.00	1.00	-0.12	0.07
coshour	0.00	0.00	0.00	0.00	0.00	-0.12	1.00	-0.10
temp	0.82	-0.29	0.03	0.00	-0.04	0.07	-0.10	1.00
hum	-0.43	-0.18	-0.01	0.01	0.00	-0.18	0.23	-0.36
calls	-0.08	0.19	-0.02	0.10	-0.07	0.50	-0.33	-0.03
rcalls	-0.09	0.18	0.01	0.08	-0.09	0.53	-0.32	-0.04
gprs	-0.13	0.15	-0.04	0.04	0.00	0.16	-0.28	-0.11
sms	-0.12	0.14	-0.16	0.18	0.03	0.11	-0.33	-0.08
adults	-0.14	0.44	-0.11	0.21	-0.07	0.00	0.00	-0.26
childs	-0.27	0.42	-0.08	0.09	0.01	0.00	0.00	-0.35
restaurant	0.14	-0.01	-0.01	0.00	0.01	0.77	-0.37	0.24
traffic	-0.12	0.41	-0.27	0.34	0.02	0.00	0.00	-0.23

*These are binary variables and correlation between them and other nominal variables was calculated using point-biserial correlation coefficient.

(continue)

	hum	calls	rcalls	gprs	sms	adults	childs	restaurant	traffic
ydsin	-0.43	-0.08	-0.09	-0.13	-0.12	-0.14	-0.27	0.14	-0.12
ydcos	-0.18	0.19	0.18	0.15	0.14	0.44	0.42	-0.01	0.41
holnum1*	-0.01	-0.02	0.01	-0.04	-0.16	-0.11	-0.08	-0.01	-0.27
holnum2*	0.01	0.10	0.08	0.04	0.18	0.21	0.09	0.00	0.34
holnum3*	0.00	-0.07	-0.09	0.00	0.03	-0.07	0.01	0.01	0.02
sinhour	-0.18	0.50	0.53	0.16	0.11	0.00	0.00	0.77	0.00
coshour	0.23	-0.33	-0.32	-0.28	-0.33	0.00	0.00	-0.37	0.00
temp	-0.36	-0.03	-0.04	-0.11	-0.08	-0.26	-0.35	0.24	-0.23
hum	1.00	-0.15	-0.15	-0.02	-0.05	0.06	0.10	-0.31	0.08
calls	-0.15	1.00	0.96	0.41	0.68	0.43	0.32	0.61	0.39
rcalls	-0.15	0.96	1.00	0.41	0.64	0.41	0.31	0.63	0.35
gprs	-0.02	0.41	0.41	1.00	0.42	0.29	0.32	0.25	0.21
sms	-0.05	0.68	0.64	0.42	1.00	0.41	0.32	0.26	0.33
adults	0.06	0.43	0.41	0.29	0.41	1.00	0.81	0.08	0.78
childs	0.10	0.32	0.31	0.32	0.32	0.81	1.00	0.03	0.58
restaurant	-0.31	0.61	0.63	0.25	0.26	0.08	0.03	1.00	0.06
traffic	0.08	0.39	0.35	0.21	0.33	0.78	0.58	0.06	1.00

environmental sciences have been discussed and described in more detail in papers by Kolehmainen et al. (2000) and Kolehmainen et al. (2001).

After training of the SOM, the attributes of its prototype vectors was used to construct the combined regional visitor attendance level model. All the variables correlated with the actual number of visitors in the area are chosen as key variables, in this case the number of telecommunication events, number of people with booked accommodation, restaurant sales and traffic density. The scaled values for the key variables on each prototype vector were summarized and the neurons were ordered according to their sum values. The neuron which had the highest sum of scores for the key variables contains the time periods when the area reached its maximum visitor attendance level, and conversely, the neuron which had the minimum sum of scores for the key variables contains the time periods when the area had its minimum attendance level. Finally, all the neurons were fitted between the maximum and minimum values and a percentage visitor attendance level model was created.

2.5. Predicting Visitor Attendance Level

In this case, the multilayer perceptron (MLP) (Haykin, 1999; Hecht-Nielsen, 1991; Gardner and Dorling, 1998) was used to create a short-term regional visitor attendance level prediction for the next seven days. The multi-layer perceptron

(MLP) is the most commonly used type of feed-forward neural network, with a structure consisting of processing elements and connections. The processing elements, called neurons, are arranged in layers, comprising an input layer, one or more hidden layers and an output layer. The input layer serves as a buffer that distributes input signals to the next layer, which is a hidden layer. Each unit in the hidden layer sums its input processes with a transfer function and distributes the result to the output layer. It is also possible for several hidden layers to be connected in the same fashion. The units in the output layer compute their output in a similar manner. For a more thorough review, the reader is referred to Haykin (1999).

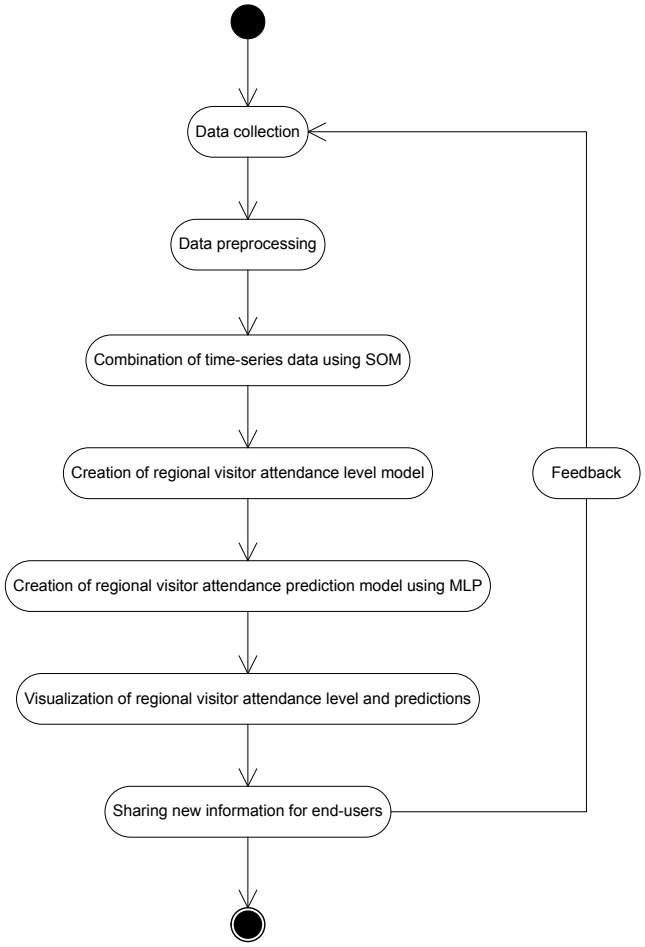


Figure 3. On-line data processing chain for predicting regional visitor attendance levels.

2.6. Visualization of Results

The results of the modelling were visualized on the form of line diagrams on various scales. As visualization with hourly resolution was found to be too detailed for long-term decision-making and planning, daily resolution was selected, and the results were similarly presented on a daily scale. The proposed automated data processing chain can be used continuously with dynamic on-line data acquisition for modelling and prediction tasks, allowing hourly and daily visitor attendance level information and short-term predictions to be fed automatically to a variety of interfaces or information-sharing systems.

2.7. Model Validation

Validation of the regional visitor attendance model was difficult, because normally there are no measuring instruments producing reliable data on the numbers of visitors in an whole area, i.e. visitor attendance levels, and where such instruments do exist, they are generally accurate only if located properly relative to the movement of visitors. This is a severe problem, because recreational areas are normally large and visitors are widely spread over them. These are main reasons for the development of new visitor monitoring and modelling technologies.

In this case, validation was carried out using expert opinions and measured freshwater use data. The annual visitor attendance level model was presented to several experts who had worked in traditional tourism companies in the Tahko area for a long time and knew activity of the area well. Moreover, the regional water use is commonly used as an indicator of visitor attendance levels as the correlation is usually good, although the problem remains that not every visitor to the area uses fresh water in the accommodation facilities or restaurants etc. Basically, water consumption describes mainly the visitors using the accommodation facilities, whereas day visitors cannot be described sufficiently accurately. In spite of these reservations, the annual visitor attendance model for the Tahko area was compared with water consumption for validation purposes.

Visualization methods such as the plotting of time-series were used here to provide an intuitive overall picture of the correlations between water consumption and visitor attendance, and statistical indices of performance were then used to validate the correspondence between the model and freshwater consumption. The numerical values of the regional visitor attendance model were on a different scale than the water use values. Therefore, values of both variables were transformed using variance scaling before comparison.

The root mean square error (RMSE), calculated according to Equation 1, was used to indicate how much of the variability in water consumption is accounted for by the visitor attendance model. RMSE is calculated according to Equation 1:

$$RMSE = \left(\frac{1}{N} \sum_{i=1}^{N} \left[P_i - O_i \right]^2 \right)^{\frac{1}{2}} \quad (1)$$

where N is the number of data points, O_i the observed data point and P_i the predicted data point. The RMSE can be divided into systematic and unsystematic components by first fitting a line by least-squares regression and then decomposing the RMSE using Equations 2 and 3:

$$RMSE_s = \left(\frac{1}{N} \sum_{i=1}^{N} \left(\hat{P}_i - O_i \right)^2 \right)^{\frac{1}{2}} \quad (2)$$

$$RMSE_u = \left(\frac{1}{N} \sum_{i=1}^{N} \left(\hat{P}_i - P_i \right)^2 \right)^{\frac{1}{2}} \quad (3)$$

where $\hat{P}_i$ is a least-squares estimate. $RMSE_s$ (systematic) describes the part of the error due to the model (linear bias). Thus, a low value implies a good model. $RMSE_u$ (unsystematic) describes the part of the error which is due to random noise and cannot be captured by the model. The index of agreement (IA), calculated using Equation 4, is a dimensionless relative measure limited to the range 0…1 which is ideal for making comparisons between models:

$$IA = 1 - \frac{\sum_{i=1}^{N} \left(P_i - O_i \right)^2}{\sum_{i=1}^{N} \left(\left| P_i' \right| + \left| O_i' \right| \right)^2} \quad (4)$$

where $P_i' = P_i - \overline{O}$ and $O_i' = O_i - \overline{O}$. The coefficient of determination (R^2), which indicates how much of the observed variability is accounted for by the model, is calculated according to Equation 5:

$$R^2 = \frac{\sum_{i=1}^{N} \left[P_i - \overline{O} \right]^2}{\sum_{i=1}^{N} \left[O_i - \overline{O} \right]^2} \quad (5)$$

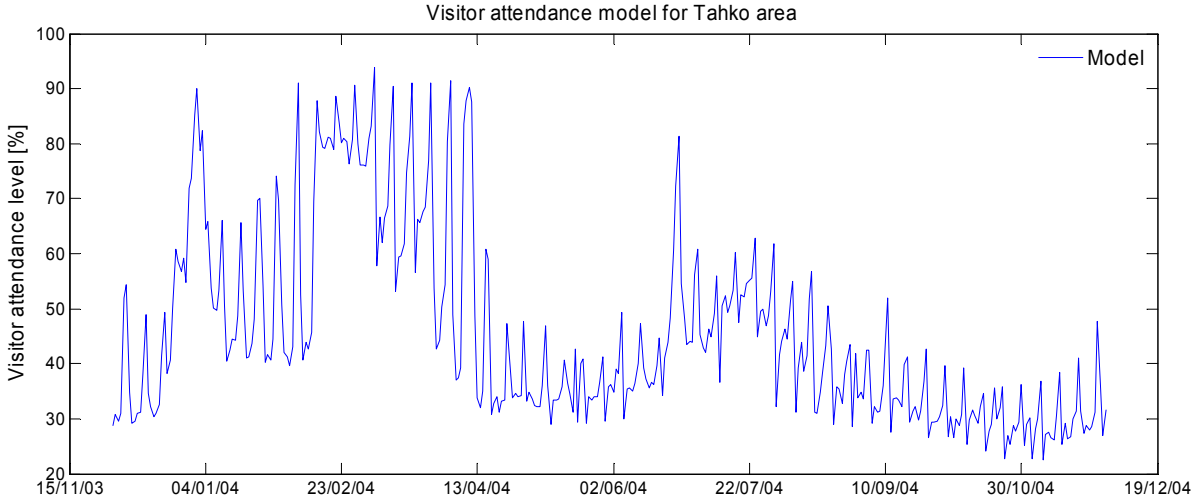
where $\overline{O}$ is the average of the observed data. The statistical indices of performance presented here are discussed in more detail by Willmot (1982) and Willmot et al. (1985). Additionally, the accuracies of the statistical indicators were investigated by calculating the standard errors (SE) using the bootstrap method (Efron and Tibshirany, 1993) with 1,000 bootstrap samples. SE is derived as a standard deviation of performance analysis.

3. Results

In this study, inferred data was used to create the regional visitor attendance level model. Information on the properties of individual variable can be valuable for regional actors and companies, as it can be used to plan business, marketing and environmental protection measures, but the most valuable information is obtained when a combined model is created. The daily visitor attendance model for the Tahko area, presented in Figure 4, shows clearly that the winter time is the main tourist season and that the weekends and national holiday periods are the most crowded times.

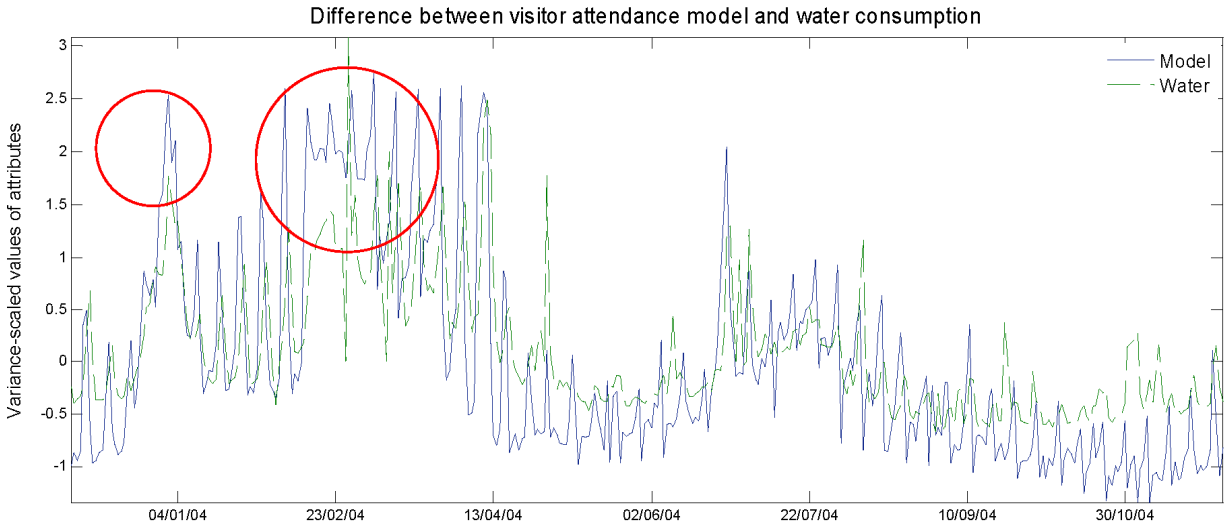
The regional visitor attendance model was validated using expert opinions and by comparison with freshwater consumption data. All the experts' opinions showed that the annual model corresponds to the real situation. One interesting point was that some of the experts commented that the difference in attendance level between winter and summer is too small, and some that the difference between weekends and working days might also is too small. Comparison with regional freshwater consumption data is presented in Figure 5 and by means of statistical indices in Table 3.

The visual comparisons showed that the variations between the seasons are very similar and the difference between weekends and working days can be seen easily in both time-series (Figure 5). Table 3 illustrates that index of agreement (IA) is 0.81 and coefficient of determination (R^2) is 0.57 which



Notes: The model shows that most crowded periods are during national holidays in the winter time, and that the summer time has a lower overall level of visitor attendance. The differences between weekends and working days can be also seen clearly.

Figure 4. Daily visitor attendance level model for the Tahko area (from 1st Dec. 2003 to 30th Nov. 2004).



Notes: The figure also shows that the main differences between visitor attendance levels and water consumption occur at the most crowded times (circled), such as New Year's Eve and national holidays.

Figure 5. Comparison of the visitor attendance level model for the Tahko area (continuous line) with water consumption (dashed line) over a period of one year (from 1st Dec. 2003 to 30th Nov. 2004) with both time-series being adjusted by variance scaling.

indicates that most of the variability in water consumption is accounted also in the created model.

The MLP model was used to perform short-term predictions of visitor attendance levels for weekly periods. The continuous system produced forecast of next seven days every morning at seven o'clock. The results were promising and daily time-series comparison between model (solid line) and predicted visitor attendance level (dashed line) is presented in Figure 6. This figure applies to the time period between 5th March and 23rd October 2005. The goodness of MLP prediction model was validated using statistical indicators and the results of calculations are shown in Table 4. Moreover, the standard errors (SE) of the indicators were calculated by using the bootstrap method with 1000 bootstrap samples and achieved results are included in Table 4.

Table 3. Statistical Indices for Model Validation Using Regional Fresh Water Consumption Data

Indicator	Value
Root mean square error (RMSE)	0.67
$RMSE_s$	0.16
$RMSE_u$	0.65
Index of agreement (IA)	0.81
Coefficient of determination (R^2)	0.57

4. Discussion

Efficient and sustainable management of tourist centres demands, not only detailed and comprehensive information about characteristics of visitors and activities, but also up to date information about number of visitors in certain time period. In this study, multi-phased data processing chain was constructed for regional visitor attendance level modelling. The method was closely related to the data sources and the results are highly dependent on the data used. The self-organizing map technique was used in a novel way to construct the combined regional visitor attendance model and multilayer perceptron was used to create short-term predictions.

The mobile telecommunications data was used in novel way, as this time-series information on the numbers of customer telecommunication events in the region was used to locate a mass of subscribers and to estimate the numbers of visitors present in the area. Mobile phone data are very valuable for this purpose because they make it possible to take account of day visitors more accurately in the attendance model. Freshwater consumption is often used as a visitor attendance indicator, but it only takes account of visitors using accommodation and restaurant services, whereas day visitors are also potential customers for many other services and affect the management and commercial development of the region and its natural environment. All types of visitors should be considered in attendance models, and this was one of the main reasons for developing the present method. In addition, the use of cell-based mobile telecommunication data makes it possible to generalize the method to other areas.

The presented results showed that individual sets of inferred data can be used to construct a combined SOM-based regional model to describe visitor attendance levels that easily detects seasonal differences and produces valuable information for end-users. On the other hand, validation of the modelling results was difficult because there is no appropriate basis for validation which describes regional visitor attendance levels or the total numbers of visitors. Expert opinions and other

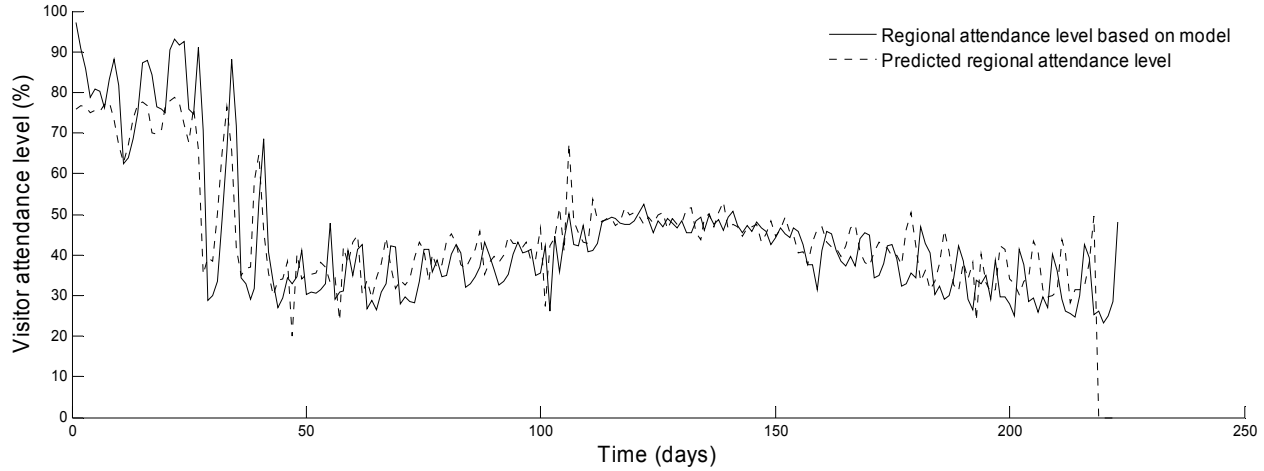


Figure 6. Regional visitor attendance level model (solid line) versus predicted situation (dashed line) for Tahko area. Resolution of the data was 24 hours and used time period was 5th March to 23th October 2005.

Table 4. The Goodness of the Predicted Visitor Attendance Model by Statistical Indicators with Bootstrap Estimates of Standard Errors

Statistical indicator	+1 day	+2 day	+3 day	+4 day	+5 day	+6 day
Index of agreement	0.92 ± 0.005	0.89 ± 0.006	0.86 ± 0.007	0.85 ± 0.007	0.87 ± 0.006	0.89 ± 0.006
Root mean square error	11.59 ± 0.29	13.23 ± 0.28	14.78 ± 0.31	15.24 ± 0.30	14.27 ± 0.30	13.14 ± 0.32
Coefficient of determination	0.70 ± 0.01	0.62 ± 0.02	0.54 ± 0.02	0.52 ± 0.02	0.57 ± 0.02	0.63 ± 0.02

comparable data therefore constituted the firmest basis for validation. The regional fresh water consumption is commonly used as an indicator of visitor attendance. For that reason, regional freshwater consumption data together with expert opinions were used to validate models and predictions. Figure 5 points to times when differences exist between the visitor attendance level and freshwater consumption, mainly in the winter and generally at the most crowded times. It is thus possible that our model based on mobile phone communication and other inferred data observes daily visitors in a more accurate way.

All the experts' opinions showed that the predictions and annual model corresponds to the real situation. Although, the interest of the companies in the tourist centre were more focused of annual models for daily visitor amount than short-term predictions. In general, the results of the validation indicate that the method developed here produces reliable information on visitor attendance levels at least in the case of the Tahko area.

Presented approach can be used continuously with dynamic on-line data acquisition for modelling and prediction tasks. In this way visitor attendance level information and predictions are always available to support different tasks of decision making. During this research project, the actors of Tahko area were using the web-based monitoring application to decide grocery shop opening periods, optimize number of operating taxi vehicles and plan more dynamic waste management in order to keep area clean as possible. This work clearly proved, however, that inferred data can be used for monitoring visitor numbers and that intelligent computational methods such as the self-organizing map are useful for improving decision-making and business planning in the field of tourism.

The presented predictive approach should be seen as a part of the bigger information system which could be used also in many other application fields than tourism. This kind of intelligent system which produces continuous near real-time information about number of people in certain place has great potential for example in air quality exposure detection, traffic census, urban town planning, disaster or emergency risk analysis or planning routing of hazardous cargo. Furthermore, living in an urban environment affects citizen health and quality of life in many ways. Moreover, in environmental impact or risk assessments, it is crucial to notify in more detailed level how many citizen are actually exposed possible undesirable phenomenons like terrible air quality or other emissions. Such information, could be linked to other environmental modeling systems like OSCAR (Sokhi et al., 2008), EXPAND (Kousa et al., 2002) or AirQUIS (Slordal et al., 2008). Overall, modern technology and data based approaches enables to calculate number of people in certain place using even smaller than hour time resolution and this kind of technology has great potential in creation of sustainable services.

5. Conclusions

The continuous predictive approach for visitor monitoring was presented and applied to model visitor attendance le-

vels in the area of Tahko. The results showed that it is possible to create reliable model for regional visitor amount by processing inferred data using presented computationally intelligent methods. Presented online monitoring system is basically tool for management and long-term planning enabling sustainable and more eco-efficient decision making in the field of tourism. Furthermore, results of this study showed that mobile telecommunication networks are useful in order to locate masses of people. These findings can bring new insight also into environmentally related problems, like risk analysis, where exposure of people needs to be evaluated. Using the presented method dynamical (time varying) distribution of number of people near exposure source can be calculated.

Acknowledgments. This research was funded by Tekes (the Technology Development Centre of Finland) and Finnet-liitto ry. We would also like to thank Mr. Tapio Halkola (Finnet-liitto ry), Mr. Jarkko Utriainen (DNA Finland Ltd.), Mr. Jari K. Mathalt (Tahkovuori Ltd), Mr. Osmo Mensalo (Finnish Road Administrator), Mr. Aku Venäläinen (Tahko Chalet) and Mr. Mikko Lehto (City of Nilsiä).

References

Brandenburg, C. and Ploner, A. (2002). Models to Predict Visitor Attendance Levels and the Presence of Specific User Groups, *Proc. of the First International Conference on Monitoring and Management of Visitor Flows in Recreational and Protected Areas*, Vienna, 166-172.

Cessford, G., Cockburn, S. and Douglas, M. (2002). Developing New Visitor Counters and their Applications for Management, *Proc. of the First International Conference on Monitoring and Management of Visitor Flows in Recreational and Protected Areas*, Vienna, 14-20.

Cottrel, S. (2002). Predictive Model of Responsible Environmental Behaviour: Application as a Visitor-Monitoring Tool, *Proc. of the First International Conference on Monitoring and Management of Visitor Flows in Recreational and Protected Areas*, Vienna, 129-135.

Drane, C., Macnaughtan, M. and Scott, C. (1998). Positioning GSM Telephones, *IEEE Communications Magazine*, 36, 46-54, doi:10.1109/35.667413.

Eagles, P., Bowman, M.E. and Tao, T. (2001). Guidelines for Tourism in Parks and Protected Areas of East Asia, *IUCN-The World Conservation Union*.

Efron, B., and Tibshirany, R. (1993). *An Introduction to the Bootstrap*, Chapman & Hall, London.

Gardner, M.W., and Dorling, S.R. (1998). Artificial neural networks (the multi-layer perceptron) - a review of applications in the atmosmospheric sciences, *Atmos. Environ.*, 32, 2627-2636, doi:10.1016/S1352-2310(97)00447-0.

González, M.C., Hidalgo, C.A., and Barabási, A-L. (2008). Understanding individual human mobility patterns, *Nature*, 453, 779-782, doi:10.1038/nature06958.

Green, D.G., and Klomp, N.I. (1998). Environmental Informatics - A new paradigm for coping with complexity in nature, *Complexity International*, 6.

Hand, D., Mannila, H., and Smyth, P. (2001). *Principles of Data Mining*, MIT Press.

Haykin, S. (1999). *Neural Networks: A Comprehensive Foundation*, Prentice-Hall.

Hecth-Nielsen, R. (1991). *Neurocomputing*, Reprinted with corrections, Addison-Wesley.

Helsinki University of Technology, Laboratory of Computer and Information Science (CIS) (2008). Home page of SOM Toolbox (Finland). http://www.cis.hut.fi/projects/somtoolbox/ (accessed March 1,2008).

Heywood, J. (1993). Behavioral Conventations in Higher Density, Day Use Wild land/Urban Recreation Settings: A Preliminary Case Study, *Journal of Leisure Research*, 25. 39-52.

Hornback, K.E., and Eagles, P. (1999). *Guidelines for public use measurement and reporting at parks and protected areas*, IUCN, Gland, Switzerland and Cambridge, UK.

Kohonen, T. (1997). *Self-Organizing Maps*, 2nd Edition, Springer-Verlag.

Kolehmainen, M. (2004). *Data exploration with self-organizing maps in environmental informatics and bioinformatics*, Ph.D. Dissertation, Laboratory of Computer and Information Science, Department of Computer Science and Engineering, Helsinki University of Technology, Helsinki, Finland.

Kolehmainen, M., Martikainen, H., Hiltunen, T., and Ruuskanen, J. (2000). Forecasting air quality parameters using hybrid neural network modeling, *Environ. Monit. Assess.*, 65, 277-286.

Kolehmainen, M., Martikainen, H., and Ruuskanen, J. (2001). Neural networks and periodic components used in air quality forecasting, *Atmos. Environ.*, 35, 815-825.

Kousa, A., Kukkonen, J., Karppinen, A., Aarnio, P, and Koskentalo, T. (2002). A model for evaluating the population exposure to ambient air pollution in an urban area, *Atmos. Environ.*, 36, 2109-2119.

Krämer, A., and Roth, R. (2002). Spatial Requirements of Outdoor Sports in the Nature Park Southern Blackforest - GIS-based Conflict Analysis and Solutions for Visitor Flow Management, *Proc. of the First International Conference on Monitoring and Management of Visitor Flows in Recreational and Protected Areas*, Vienna, 33-39.

Lim, C., and McAleer, M. (2005). Ecologically sustainable tourism management, *Environ. Model. Software*, 20, 1431-1438.

The Mathworks Inc. Home page (2008). http://www.mathworks.com/ (accessed March 1, 2008).

Muhar, A., Arnberger, A., and Brandenburg, C. (2002). Methods for Visitor Monitoring in Recreational and Protected Areas: An Overview, *Proc. of the First International Conference on Monitoring and Management of Visitor Flows in Recreational and Protected Areas*, Vienna, 1-6.

Niska, H., Hiltunen, T., Karppinen, A., Ruuskanen, J., and Kolehmainen, M. (2004). Evolving the neural network model for forecasting air pollution time series, *Engineering Applications of Artificial Intelligence*, 17, 159-167.

Page, B., and Raustenstrauch, C. (2001). *Introduction to environmental information systems*, Environmental information systems in industry and public administration, Idea Group Publishing.

Santasalo, T. (2007). *Matkailukohteiden kävijämäärät 2006*, Matkailun edistämiskeskus.

Slordal, L.H., Mc Innes, H., and Krognes, T. (2008). The Air Quality Information System AirQUIS, *Information Technologies in Environmental Engineering*, Special Issue. January 2008.

Sokhi, R., Mao, H., Srimath, S., Fan, S., Kitwiroon, N., Luhana, L., Kukkonen, J., Haakana, M., Karppinen, A., Dick van den Hout, K., Boulter, P., McCrae, I., Larssen, S., Gjerstad, K., San, José R., Bartzis, J., Neofytou, P., van den Breemer, P., Neville, S., Kousa, A., Cortes, B., and Myrtveit, I. (2008). An integrated multi-model approach for air quality assessment: Development and evaluation of the OSCAR Air Quality Assessment System, *Environ. Model. Software*, 23, 268-281.

Statistics Finland (2007). Infrastruktuuri, matkaviestintä: Matkapuhelinliittymien määrä sekä liittymät 100 asukasta kohti vuosina 1980, 1985 ja 1990-2005 (Finland), http://www.stat.fi/til/tvie/2005/tvie_

2005_2006-06-01_tau_009.html (accessed March 1, 2008).

Tahkovuori Ltd. (2005). Company's home page (Finland). http://www.tahko.fi/index.php?cId=45 (accessed March 1, 2008).

Ultsch, A., and Siemon, P. (1990). Kohonen's Self-Organising Feature Maps for Exploratory Data Analysis, *Proc. of the International Neural Network Conference (INN'90)*, Dordrecht, 305-308.

Willmot, C. (1982). Some comments on the evaluation of the model performance, *Bulletin of American Meteorological Society*, 63, 1309-1313, doi:10.1175/1520-0477(1982)063<1309:SCOTEO>2.0.CO;2.

Willmot, C., Ackleson, S., Davis, R., Feddema, J., Klink, K., Legates, D, O'Donnell, J., and Rowe, C. (1985). Statistics for evaluation and comparison of models, *J. Geophy. Res.*, 90, 8995-9005, doi:10.1029/JC090iC05p08995.