



A machine-learning algorithm for wind gust prediction

P.J. Sallis^{a,*}, W. Cluster^b, S. Hernández^c

^a Geoinformatics Research Centre, Auckland University of Technology, Auckland, New Zealand

^b Environmental Research Laboratory, Ritsumeikan Asia Pacific University, Beppu, Japan

^c Laboratorio de Procesamiento de Información Geoespacial, Universidad Católica del Maule, Talca, Chile

ARTICLE INFO

Article history:

Received 11 February 2010

Received in revised form

9 November 2010

Accepted 11 November 2010

Available online 24 March 2011

Keywords:

Wind velocity modeling

Wind gust prediction

Machine-learning algorithms

Geostatistics

ABSTRACT

Physical damage to property and crops caused by unanticipated wind gusts is a well understood phenomenon. Predicting its occurrence continues to be a challenge for meteorologists and climatologists. Various approaches to gust occurrence model building have been proposed. The very nature of the event is problematic because of its brief duration following a rapid change of state in wind velocity that immediately precedes it. Events classified as wind gusts have a typical duration of less than 20 s and are often much shorter. The rapidly accelerating wind velocity preceding them is often not apparent until the gust occurs. They come quickly, occur suddenly, and then end as abruptly as they began.

Observations of 2000 gust events were made during the research to which this paper refers. These observations indicated a mean interval of 3.2 min between the beginning and end of wind velocity change and a noticeable linear progression in the acceleration pattern. It was also noted that state changes regularly occur, often over only seconds in time. In combination, these factors pose both a sampling and a data interpretation challenge, making reliable prediction difficult. This paper describes some new research undertaken to investigate methods of wind gust measurement and prediction. In particular, a machine-learning approach is taken to determine a satisfactory analytical process and to produce meaningful and useful results. An algorithm for use with real-time climate data collection and analysis is proposed, with a description of its implementation. Real-time data sampling provides input for this study using terrestrial sensor telemetry. Near-ground truth data are recorded independent of geostrophic upper atmosphere conditions.

© 2011 Elsevier Ltd. All rights reserved.

1. Introduction

The term prediction, when formally defined, may not be the best way to consider the expectation of a future wind gust event. Forecasting and prediction are concerned with estimating the state of a dynamical system using historical data. Instead, anticipation can be achieved by analyzing several factors present at a point in time, so that we can give an estimate of the likelihood of a gust event for the next period. In this paper, we propose an operational prediction system based on observed climatic conditions using near-surface half-hourly averaged observations.

Gusts, if considered as intrinsically unpredictable events, are complex in nature because of the complicated superposition of explanatory variables. Defining these variables and their relationships is nontrivial. Exposing these for processing by a machine-learning algorithm such that wind events may be reliably anticipated is at the heart of the work described in this paper.

Numerous studies relating to the wind gust problem have been carried out using deterministic, probabilistic, and stochastic models, but none conclusively indicates a single robust method for predicting gust events (Sfetsos, 2000; Alexiadis et al., 1998). Work reported in Bierbooms (2005) utilizes a constrained stochastic simulation introducing a wind turbulence variable for discrete events (similar, it is said, to the autocorrelation function of turbulence). This study demonstrated that for wind turbines, the generation of stochastic time series was useful for producing reliable results. In another study where computational neural networks were used to predict wind speed, it was noted that existing forecasting systems are capable of providing realistic atmosphere dynamic approximations (Mohandes et al., 1998). They showed that to ensure the stability of the numerical methods used, a data assimilation process is needed to represent what is considered a sparse and noisy set of observations. According to U.S. weather observation practice, gusts are reported when the peak wind speed (pws) reaches at least 16 knots and the variation in wind speed (vws) between the peaks and lulls is at least 9 knots (see Cook and Gruenbacher, 2008). If we add to these two variables pws and vws the mean velocity acceleration (mva)

* Corresponding author.

E-mail address: psallis@aut.ac.nz (P.J. Sallis).

of say 3.2 min and 0 gust duration (gd) of between 1 and 20 s, we can consider a complex yet interdependent variable set for the prediction of a gust event (where $gp=1$ or “true”) with the Boolean expression form

```
(if pws >= 16 and vws >= 9) and
(if mva = 3.2 and gd > 1 or gd < 20) then gp = 1
```

This expression can be considered suitable for algorithmic implementation where real-time data are processed as a constant stream input.

The physical considerations that lead to wind gusts have been studied, describing a wind gust estimate as a deterministic function of the atmospheric turbulence and wind speed in the boundary layer (Brasseur, 2001). Wind gusts, as with other natural phenomena, are usually regarded as unpredictable. Therefore, we consider that for any model of these, some randomizing element should be incorporated for the unexplained variation. For this reason, it was felt sufficient for this study to model near-ground truth data in an endeavor to prove the prediction algorithm. Moreover, it has recently been noted that the accuracy of the physical model is highly dependent on the quality of the simulated flow (Agústsson and Ólafsson, 2009). Further research has concentrated on factor derivation for gusts based on wind speed state changes with acceleration and deceleration derivatives over time ($v=dx/dt$), and in some cases, as for recent research in wind turbine placement research, this has been shown to be best treated as a stochastic rather than deterministic process (Lei et al., 2009).

Using an alternative methodology, other recent research has used microwave radiometer data to improve the so-called gust expectancy (GUSTEX) by using empirical information provided twice daily (Chan and Wong, 2008). The authors proposed this model to closely monitor these mostly unpredictable events, especially during the seasons of spring and summer, when they are most prevalent. The prediction method relies almost entirely on wind velocity acceleration and mean peak force against deceleration over time. Diurnal and seasonal variations are necessary algorithmic factors, which implies that any analytical method should monitor data over long periods of time and certainly at least over a four-season period.

In this paper, a new approach to data analysis and problem solution based on empirical relationships is proposed. Some a priori assumptions relating near-surface air pressure, temperature, and humidity as wind gust event factors were tested using real-time data gathered from a global network of climate monitoring stations. This network of instrument sets is being used for micro-climate monitoring and modeling in vineyards and is the subject of a wider research program participated in by the authors in Sallis et al. (2008). Although these data comes from a relatively short time period (2 y at best), they do indicate that the a priori assumptions produce a reliable classification dependency set (0.76) when processed using a multilayer perceptron (MLP) algorithm implemented by Weka,¹ a published data mining product. Running experiments with a linear regression algorithm also produced a convincing result, sufficient to determine for the researchers that further work could result from the construction of a new transform-function-based algorithm. This points to the potential development of a machine-learning solution for predicting wind gusts that could use a richer data set than those available in the past, including other variables such as air temperature and humidity. This is considered a unique approach to solving the wind gust problem, yet one that requires significant experimentation with large data sets in order to prove its

veracity. This potential result is supported by previous studies (e.g., Kretzschmar et al., 2004; Gneiting et al., 2006), where the investigation of problems with similar characteristics led to the conclusion that more variables were needed to adequately model the complexities of random events in nature.

The research methodology described below is a procedure for preprocessing sample data prior to applying the analytical algorithms chosen for testing. The initial data exploration process and visualization of results are also outlined. A brief description of the models tested and their results is given. Finally, the implementation of this methodology for use in an operational real-time monitoring system is presented with a discussion of conclusions from the study.

2. Research methodology

For the purposes of this research, several algorithms were examined for goodness of fit. These algorithms are described in Section 7. The algorithms were tested using continuous stream sampling from six climate data collection locations in three countries. The purpose of this sample structure is to demonstrate consistency of analysis and results from disparate geographical locations, and therefore the potential generalizability of the method. In the first instance, to establish the analytical framework for the study, a single location was used with data sets of seven attributes, which are examined with a total of 1906 records recorded daily in a Northern New Zealand Winery at Kumeu River <http://www.kumeuriver.co.nz>. In the second study both this site and a second site in Chile (Casa Donoso vineyard <http://www.casadonoso.cl>) were used with the same seven variables. Based on assumptions that the relationship between those seven variables would lead to a prediction of gust events in the coming days, a predictive model was built using decision trees, logistic regression, and multilayer perceptron (MLP) neural network algorithms. These algorithms were evaluated for their performance using a comparative accuracy matrix algorithm and the results from this analysis are discussed.

3. Description of the data

The initial single-location data set for this study records seven meteorological attributes. The input features selected for the classifiers were hour of day, temperature, humidity, rainfall, pressure, wind speed, and dew point. Dew point is a derived value but is necessary to establish a value point in this data set. Analysis of the data from the Kumeu weather station was carried out for three 40-day periods in February, April, and June of 2009. The attributes are time-dependent and are recorded daily every half hour. The gust variable, although recorded as numeric, was transformed into a binary variable indicating the existence or absence of a gust event. Initial tests with 1-min and 1-h sampling intervals suggested either too short or too long a period for prediction, based on the 3.2-min event separation average determined in early observations. Given that not every wind velocity acceleration resulted in a gust, it was decided to sample at 30-min intervals from the continuous data recorded by the sensors.

We then conducted a second set of tests where the three best performing models were evaluated on two 90-day data sets; one from Casa Donoso in the Maule Region of Chile and one from Kumeu River in Northern New Zealand. By examining the relationship among these attributes, an algorithm can be constructed that provides as output indications of the likelihood of a gust event.

¹ <http://www.cs.waikato.ac.nz/ml/weka/index.html>

4. Data preprocessing

In the first experiment there are a total of seven input attributes with 1906 records in the data set. These records have been separated into two sets labeled as training and testing sets, with the first 66% (1258 records) of the records in the training set and the last 34% in the test set. All variables other than Gust are numeric. As mentioned above, Gust was transformed to a 0/1 nominal variable (0, no gust; 1, yes gust). Date was ignored but time of day was included.

The second experiment consisted of similar data, but collected over a 90-day period beginning June 22, 2010. Two sets of data were from separate locations. There are no missing data values in the data sets. There are a few outliers, which may be the result of extreme weather patterns on certain hours of some days. These are accounted for in the analysis and the results described in Section 8.

5. Data exploration and visualization

The following will refer to the April data from the Kumeu site. The data for February and June will not be described here, but the remarks extend to these months without significant deviation. The histogram of each attribute was visualized to reflect their distributions and class positions in each attribute, as shown in Fig. 1. False represents class 0, no gust, and True is class 1, yes gust.

When the histograms in Fig. 1, are observed, none of the distributions appear to be Gaussian, and this was confirmed by conducting a normality test. Data values for temperature appear not to be too far from a normal distribution, although application of the one-sample Kolmogorov–Smirnov test for normality shows that actually a normal distribution is not a good fit to the data. This result might be expected, because these parameters are driven by diurnal and seasonal variations.

Gust events can be seen to begin when the temperature is approximately 7 °C. In other words, a temperature around 7 °C

and above may indicate the potential occurrence of a gust event. Conversely, when a lower temperature prevails, a gust event is unlikely. Table 1 shows the maximum, minimum, mean, and standard deviation for the first experiment data.

The humidity histogram in Fig. 1b shows both the data and class distribution of the humidity. The histogram is skewed to the left, indicating that large portion of data are associated with higher humidity. The occurrence of class 1 can be seen throughout the distribution. The histogram for pressure is skewed to the left, as almost all the pressure falls below 1030.4 hp. Very few records are above 1039.4, and there are outliers in the data set (see Fig. 1d). There is no clear distinction between classes 0 and 1 according to changes in atmospheric pressure.

The histogram for the wind speed variable is skewed to the left with a large portion of data exhibiting low wind speed (see Fig. 1f). According to Kretzschmar (2002), the skew between weak and severe winds is on the order of 100:1. We can observe here that there are a few outliers in this variable. Moreover, the histogram also indicates that the occurrence of class 1 increases as wind speed increases. By examining scatterplots in Fig. 2, relationships between pairs of variables are observed. It appears that in a specific range of temperature and wind speed values, occurrences of wind gust events are likely.

On further analysis, though, we note a peculiar relationship between future gust events (30 min in the future) and current

Table 1
Basic statistics for data in the training set.

	Temperature	Humidity	Rainfall	Pressure	Wind speed	Dew point
Mean	10.09	82.56	49.53	1010.51	4.31	7.13
Median	11.00	86.00	30.00	1011.80	0.00	7.70
Mode	11.70	91.00	42.40	1011.80	0.00	7.10
Std. deviation	3.88	9.90	178.98	9.66	7.75	3.40
Minimum	-1.70	40.00	0.00	988.10	0.00	-3.00
Maximum	16.30	94.00	1354.00	1082.70	57.20	12.90

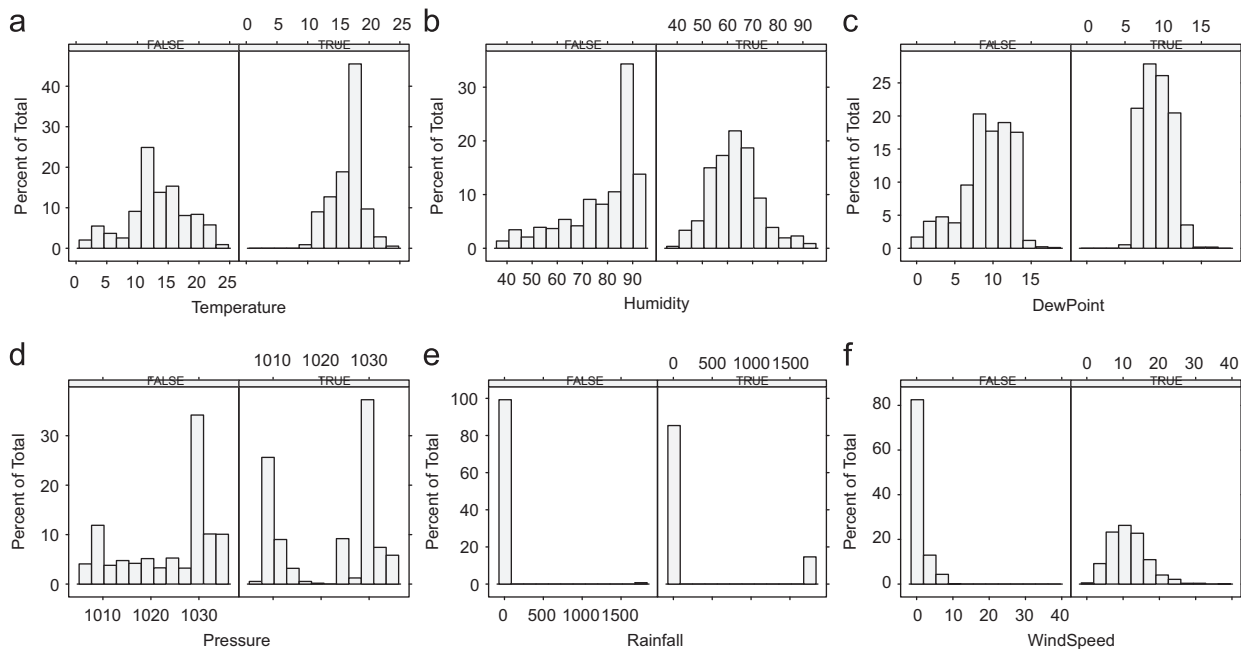


Fig. 1. Distribution of variables and classes. False indicates no gust. True indicates gust event. (a) Temperature. (b) Humidity. (c) Dew point. (d) Pressure. (e) Rainfall. (f) Wind speed.

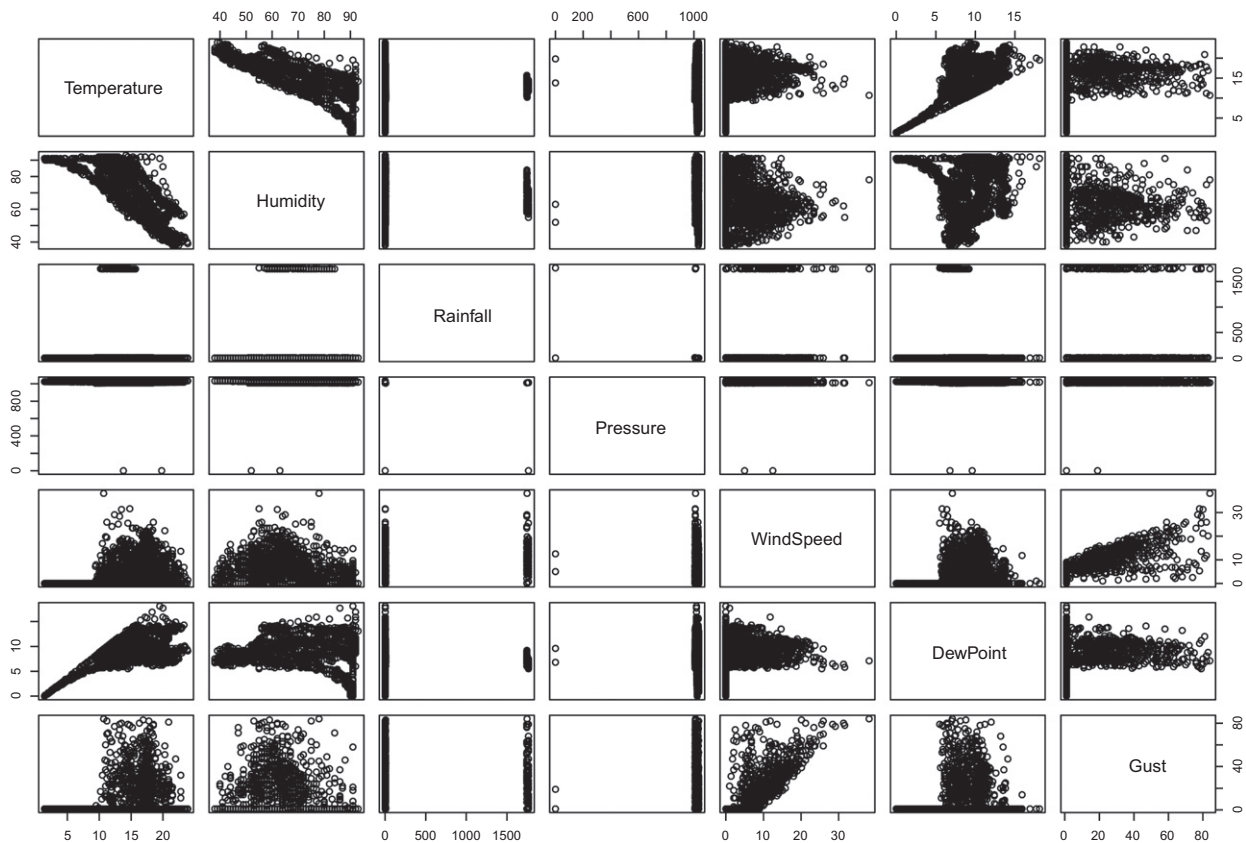


Fig. 2. Scatterplot of climate data and gust event.

wind speed in the scatterplot in Fig. 3. In the figure, the black overlay indicates no future gust event and the red overlay shows the presence of a future gust event. It can be observed that a predominance of the records without a future gust event have current low wind speeds. This leads to the conclusion that wind gust events are intrinsically the result of a complex set of variables and dependency values. The prediction of unknown values is therefore difficult. To strengthen the findings from the results represented in the histograms and scatterplots, a number of data mining algorithms are tested and compared to determine dependency relationships for the data sets.

6. Initial regression model

For initial exploration of the data a multivariate regression was used, which resulted in a multiple R value of 0.6976 and an adjusted R^2 value of 0.4866. The coefficients for this model are given in Table 2. This was considered sufficient to determine that further work could result in the construction of a new transform-function-based algorithm for predicting wind gusts if a richer data set including air temperature and humidity factors was used.

7. Models tested

7.1. Logistic regression

Logistic regression is a statistical model that is used when the outcome is binary in nature. It relates the log odds of $\Pr(\text{event})$ to a linear combination of predictor variables. In the experiments described here, a multinomial logistic regression model with a ridge estimator is used, as described by Kretzschmar (2002). This

method has been shown in our studies to be both fast and accurate for classification tasks (Le Cessie, 1992).

7.2. Neural network

An artificial neural network (also referred to here as a multi-layer perceptron or MLP) is a model used to predict outcomes based on inputs. It consists of an interconnected group of artificial neurons. Neural networks are nonlinear statistical modeling tools and can be used for supervised or unsupervised variable relationship or dependency learning (Lim and Shih, 2000; Hastie et al., 2001). Here they are used for supervised learning.

7.3. Simple logistic

This is a classifier for building linear logistic regression models. LogitBoost with simple regression functions as base learners is used for fitting the logistic model (Landwehr and Frank, 2005; Sumner et al., 2005).

7.4. The C4.5 tree (J48)

The C4.5 algorithm (Quinlan, 1993) constructs a decision tree using the concept of information entropy. At each node C4.5 chooses one covariate that most effectively separates the records into those within one of the binary classifications and those within the other binary classifications. The method only allows symbolic output fields, but C4.5 can support splits at each node that result in more than two subgroups for symbolic predictor fields.

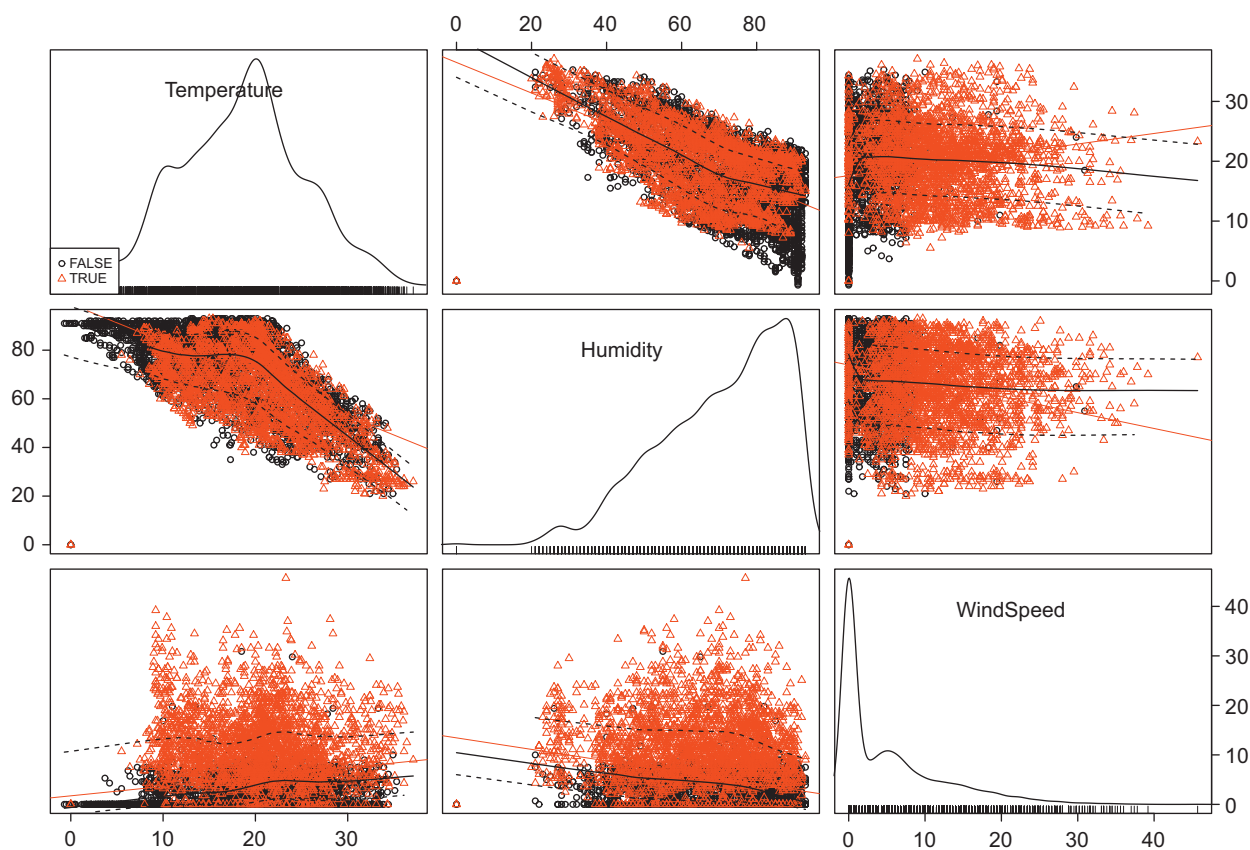


Fig. 3. Scatterplot of climate data and occurrence of a future gust event. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 2
Coefficients of the regression model.

	Coefficients	Standard error	tStat	P-value
Intercept	58.52714817	14.08631772	4.154893376	3.37E-05
Hour	0.202377739	0.023742201	8.523967039	2.67E-17
Temperature	2.194109471	0.677288728	3.239548185	0.001213
Humidity	0.620366511	0.142303147	4.359471481	1.36E-05
Rainfall	0.000975886	0.000855243	-1.1410626	0.253957
Pressure	0.002970214	0.003334494	-0.890754001	0.37315
Wind speed	0.885637433	0.032334569	27.38980188	4.5E-144
Dew point	2.388886641	0.688340246	3.470502641	0.000529

7.5. Simple CART tree

The CART (classification and regression trees) algorithm is a nonparametric technique that can handle categorical or numeric dependent variables and this distinguishes it from other decision trees methods such as C4.5. It produces binary splits at each node (Breiman et al., 1984).

8. Results from experiments

8.1. Data classification

As mentioned above, the analytical algorithms used with this data are from the data mining software suite WEKA. They are particularly appropriate when examining continuous time-dependent attribute value vectors. In particular, the implementations of classification algorithms for MLP neural networks, logistic regression, simple logistic regression, and J48 were tested. The J48

algorithm is Weka's implementation of the C4.5 decision tree as outlined in Quinlan (1993). Random Forests and Simple Cart, which is Weka's implementation of CART, are described in Breiman et al. (1984) and Breiman (2001).

The algorithms were tested on three sets of data, each of 40 days duration. Each 40-day set was structured in four ways. They were tested with and without the inclusion of the wind parameter. Additionally, the current gust event was tested, and then further with gust events occurring 30 min in the future. The sampling times for data collection at the various sites are currently set at 30-min intervals. So the following tests were conducted:

1. A data set with wind speed included in the set of input variables was used to predict the occurrence of a gust event occurring at the same instant that the input variables were recorded.
2. A data set with wind speed excluded from the set of input variables was used to predict the occurrence of a gust event occurring at the same instant that the input variables were recorded.
3. A data set with wind speed excluded from the set of input variables was used to predict the occurrence of a gust event occurring 30 min after the input variables were recorded.
4. A data set with wind speed included was used to predict the occurrence of a gust event occurring 30 min after the input variables were recorded.

Thus, there were $3 \times 4 = 12$ variations of the data. These 12 variations of the data were tested on the six classification algorithms mentioned above. The resulting 12×6 table is given in Table 3.

Table 3
Gust prediction over three 40-day periods beginning in February, April, and June of 2009. Six models tested with four different data sets.

Data sets	Neural net	Logistic	Simple logistic	J48 tree	Random forest	Simple Cart
April_NWS	75.49	73.54	73.54	76.94	73.54	77.06
April_WS	91.02	95.27	95.39	95.63	95.02	95.15
April_Lag_WS	78.88	81.43	83.13	86.77	77.06	87.01
April_Lag_NWS	75.97	70.63	71.6	72.57	74.64	75.12
Feb_NWS	85.93	85.82	85.82	78.37	72.81	64.89
Feb_WS	91.02	92.67	92.2	91.02	91.37	91.37
Feb_Lag_WS	84.99	85.82	82.39	79.31	79.91	74.7
Feb_Lag_NWS	87.83	89.72	90.07	88.53	86.29	88.42
Jun_NWS	72.27	75.44	76.39	68.46	72.27	72.27
Jun_WS	92.39	94.29	94.29	94.45	93.98	94.14
Jun_Lag_WS	83.2	83.52	83.36	79.56	81.62	82.73
Jun_Lag_NWS	74.33	77.5	76.7	72.27	70.21	72.27

Table 4
Nomenclature used in Table 3.

Description	
NWS	Predicted gust simultaneously with inputs, wind speed NOT included as input
WS	Predicted gust simultaneously with inputs, wind speed included as input
Lag_NWS	Input lagged by 30 min, wind speed NOT included as input
Lag_WS	Input lagged by 30 min, wind speed included as input

8.2. Comparison of models

Table 3 shows the results of the initial round of tests. It has already been noted that wind speed is highly correlated with the gust event. If wind speed is included, the current gust event should be predictable with a high degree of accuracy. This is confirmed in Table 3 in the data sets labeled April_WS, Feb_WS, and Jun_WS. Table 4 Describes the nomenclature used in Table 3. Although the predictions where the input data and the output data are both recorded at the same instant manifest the highest levels of accuracy, they are less useful because they appear to be unactionable. It is of course, desirable to predict future gust events rather than only current gust events. The prediction of current gust events has an average accuracy of 93.37%. The average for all three of these 40-day data sets is 84.39. SimpleCart is the best-performing model, with an accuracy of 86.05, followed by simple logistic and J48, with accuracies of 85.52 and 84.95, respectively. The relationship between gusts and wind speed, together with the other variables in the set, appears to be mildly nonlinear. This would suggest using linear regression for a better fit than with neural networks.

With these tests completed, the three best-performing models (simple logistic, J48 tree, and SimpleCart) were evaluated. These experiments used 2 × 90-day data sets; one from Casa Donoso in the Maule Region of Chile and one from Kumeu River in Northern New Zealand. This was the second round of experiments. Class imbalances were addressed by using SMOTE (synthetic minority over-sampling technique) (see Chawla et al., 2002). Table 5 shows the three models compared with a baseline of ZeroR. ZeroR is a classifier that simply predicts the mean (for a numeric class) or the mode (for a nominal class). The ZeroR model does not take into account input variables and therefore, allows for a measure of improvement of other models over chance prediction (the ZeroR predictions).

In Table 5, the bold type face indicates models with superior performance at the 5% significance level as compared with a baseline determined by the simple logistic model.

Table 5
Best-performing models compared with the baseline ZeroR model.

Data set	ZeroR	Simple logistic	Simple Cart	J48
Casa Donoso	58.25	88.79	89.30	89.60
Kumeu	55.41	87.78	87.59	87.53

Table 6
Overall accuracy for three models tested on two 90-day data sets.

Data set	Simple logistic	Simple Cart	J48
Casa Donoso	84.28 (1.08)	89.60 (1.02)	89.30 (0.88)
Kumeu	85.44 (0.89)	87.53 (0.88)	87.59 (0.86)

Table 7
False positive rates for three models tested on two 90-day data sets.

Data set	Simple logistic	Simple Cart	J48
Casa Donoso	0.12 (0.01)	0.07 (0.02)	0.08 (0.01)
Kumeu	0.11 (0.01)	0.14 (0.02)	0.13 (0.01)

Simple Cart is now used as a baseline model in Table 6, which shows that J48 and SimpleCart performed better than the baseline at the 5% significance level. Numbers in parentheses to the right of accuracy are standard deviations for the (10 × 10)-fold cross-validation tests.

Tests for FPR (false positive rate) were conducted because reduction of this measure is necessary to catch the occurrence of a gust event even at the expense of generating a false alarm. Table 7 shows the results. In terms of the false positive rate the results are not determinative, as they differ on the two data sets. On the Casa Donoso set, J48 and SimpleCart outperform simple logistic regression, but on the Kumeu data set, simple logistic performs better. In terms of average over the two data sets, simple logistic performed best, with 11.5% as compared with J48 and SimpleCart, both with 10.5% FPR. The overall average FPR for the 3 models was 10.8%.

In Table 7, the bold type face indicates models with superior performance (false positive rate is higher) at the 5% significance level as compared with a baseline determined by the simple logistic model. The numbers in parentheses to the right of accuracy are standard deviations for the (10 × 10)-fold cross-validation tests.

8.3. The J48 tree visualization

A decision tree (Fig. 4) was constructed using the J48 algorithm with both default parameters and an unfiltered data set. Among the algorithms tested, the decision tree offers the opportunity for visualizing how the classification of gust (0,1) is obtained through the attributes.

When wind speed is less than 2.8, there are no gust events predicted. There were 6308 cases predicted accurately and 178 cases incorrectly identified. If wind speed is greater than 2.8 and greater than 8.6, then a gust event is predicted (with 829 cases predicted accurately and 214 predicted incorrectly). The tree can be followed and read in this way. In summary, a gust event will occur only when one of the following sets of conditions are met. (See the red arrows on the tree).

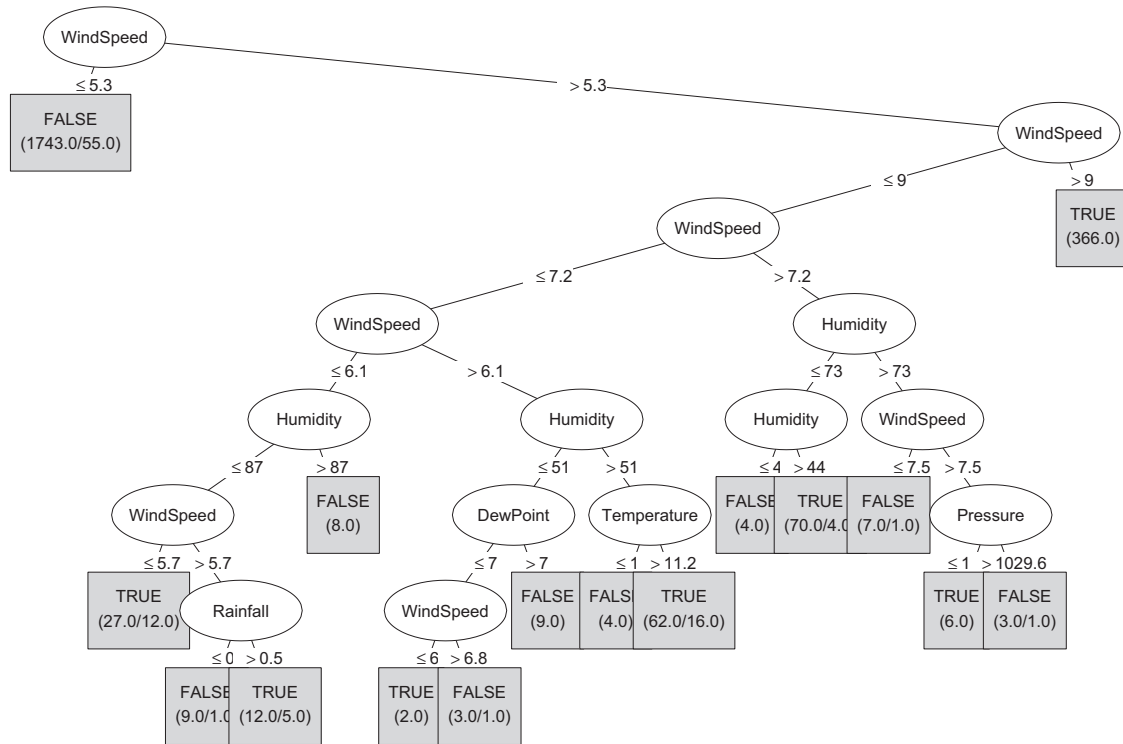


Fig. 4. J48 decision tree.

9. Binding the algorithm into an operational real-time climate monitoring system

The Simple Cart method generates a decision tree, which, when converted into a set of induction rules (Fig. 4), can be viewed as a prealgorithmic description for a computer program that can be used to monitor data from weather stations or climate sampling instruments. The program could potentially anticipate/predict wind gusts using this approach. To test this proposition, a program based on the induction rules was written in C# and tested against live data from a real-time transmission stream. The performance of this program is described below. The functionality of the program is to be incorporated into an operational live climate monitoring network system developed by Ghobakhlou et al. (2010). It was transcribed into PHP and implemented as a Web-enabled component. The results from this implementation will be formulated in future work for this research following at least 2 y of data collection. This time period is necessary for reliable testing to be performed. First the induction rules are given, and then a schematic of the real-time system into which the program is incorporated is shown as Table 8.

Numbers in parenthesis in each leaf give an estimate of success at that leaf. For example, the induction rule at the bottom of Program 8:

WindSpeed > 9 : TRUE (366.0)

indicates that 366 cases defined by this node were correctly classified as 1 (a gust event occurring).

10. Discussion and conclusions

The research described here is part of a wider environmental monitoring and modeling project. The authors and their colleagues in eight countries have designed a telemetry system and built a set of monitoring instruments (sensors); that operate in

Table 8

Example of induction rules derived from J48 decision tree.

J48 pruned tree
WindSpeed <= 5.3:FALSE (1743.0/55.0)
WindSpeed > 5.3
WindSpeed <= 9
WindSpeed <= 7.2
WindSpeed <= 6.1
Humidity <= 87
WindSpeed <= 5.7:TRUE (27.0/12.0)
WindSpeed > 5.7
Rainfall <= 0.5:FALSE (9.0/1.0)
Rainfall > 0.5:TRUE (12.0/5.0)
Humidity > 87:FALSE (8.0)
WindSpeed > 6.1
Humidity <= 51
DewPoint <= 7
WindSpeed <= 6.8:TRUE (2.0)
WindSpeed > 6.8:FALSE (3.0/1.0)
DewPoint > 7:FALSE (9.0)
Humidity > 51
Temperature <= 11.2:FALSE (4.0)
Temperature > 11.2:TRUE (62.0/16.0)
WindSpeed > 7.2
Humidity <= 73
Humidity <= 44:FALSE (4.0)
Humidity > 44:TRUE (70.0/4.0)
Humidity > 73
WindSpeed <= 7.5:FALSE (7.0/1.0)
WindSpeed > 7.5
Pressure <= 1029.6:TRUE (6.0)
Pressure > 1029.6:FALSE (3.0/1.0)
WindSpeed > 9:TRUE (366.0)
Number of Leaves: 16
Size of the tree: 31

real-time to sample climate, atmosphere, plant, and soil data. The project began with monitoring these data for microclimates in vineyards but has now been extended to orchards and other fruit and vegetable crops. These instruments have now been placed in

some 30 locations throughout the participating countries. One of the sets of data recorded for various modeling purposes such as frost prediction and crop quality is wind velocity. During the study of wind velocity, the phenomenon of wind gusts was considered to be an influential factor in plant wellness and crop quality. Also, it was observed that wind gusts were responsible in some cases for considerable damage to crops, especially in orchards. Building on the assumption that wind gusts were an important factor in microclimate monitoring and modeling, the work described here applied computational and statistical methods used in general by the authors to the wind gust problem domain, which is described elsewhere in published research. The work evaluated the algorithmic quality and appropriateness potential of six machine-learning classifiers when applied to streaming data of observed wind gust events sampled at 30-min intervals. The classifiers presented here exhibited superior quality when compared with a ZeroR model with respect to the evaluation measure of overall accuracy for the prediction of wind gust events with a lead time of 30 min. FPR (false alarm rates) for the top three performing models had a mean value of 10.8%. The input features selected for the classifiers were hour of day, temperature, humidity, rainfall, pressure, wind speed, and dew point. The dependent variable was a binary variable reflecting the existence or lack of existence of a gust event. The data were sampled from two sites currently being monitored by the research team in the telemetry grid mentioned above, a northern New Zealand winery and one of similar topographical characteristic location in the Maule region of Chile. Various techniques were used to ensure statistical robustness in the data analysis. The problem of class imbalance, for example, was addressed by employing synthetic minority oversampling technique. The mean overall accuracies for the three top performing models measured on the 90-day data sets were 84.86% for simple logistic, 88.00% for the J48 decision tree, and 88.45% for the Simple Cart decision tree. This analysis was carried out using historical time series data, and from it the J48 decision tree algorithm was selected as being the most useful for predicting, or at least anticipating, a wind gust event given the set of variable values typically observed in our sample. In order to test the algorithm with live streaming data, we developed a computer program, which was then implemented as an analytical component of our operational real-time telemetry system. The activity of this program is now being monitored for its output results, which are in the form of decisions relating to the possibility of a wind gust event occurring at any time. Because seasonal and other climate and atmospheric variations need to be examined when a wind gust event is observed and leading up to such an event, the authors intend to analyze the data set over an initial 1-y period, and then a second and third year. The results from these observations will be published if and when they prove

to be conclusive and appropriate for useful wind gust prediction or anticipation.

References

- Agústsson, H., Ólafsson, H., 2009. Forecasting wind gusts in complex terrain. *Meteorology and Atmospheric Physics* 103, 173–185.
- Alexiadis, M.C., Dokopoulos, P.S., Sahsamanoglou, H.S., Manousaridis, I.M., 1998. Short-term forecasting of wind speed and related electrical power. *Solar Energy* 63, 61–68.
- Bierbooms, W., 2005. Constrained stochastic simulation—generation of time series around some specific event in a normal process. *Extremes* 8, 207–224.
- Brasseur, O., 2001. Development and application of a physical approach to estimating wind gusts. *Monthly Weather Review* 129, 5–25.
- Breiman, L., 2001. Random forests. *Machine Learning* 45, 5–32.
- Breiman, L., Friedman, J., Olshen, R.A., Stone, C.J., 1984. *Classification and Regression Trees*. Wadsworth International Group.
- Chan, P.W., Wong, K.H., 2008. Application of microwave radiometer and wind profiler data in the estimation of wind gust associated with intense convective weather. In: *Proceedings of the IOP Conference Series: Earth and Environmental Science*.
- Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P., 2002. Smote: synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research* 16, 321–357.
- Cook, K.R., Gruenbacher, B., 2008. Assessment of methodologies to forecast, wind gust speed national weather service.
- Ghobakhlou, A., Perera, A., Sallis, P., Zandi, S., 2010. Modular sensor nodes for environmental data monitoring. In: *Proceeding of 4th International Conference on Sensing Technology (ICST)*.
- Gneiting, T., Larson, K., Westrick, K., Genton, M., Aldrich, 2006. Calibrated probabilistic forecasting at the stateline wind energy center. *Journal of the American Statistical Association* 101, 968–979.
- Hastie, T., Tibshirani, R., Friedman, J., 2001. *The Elements of Statistical Learning*. Springer.
- Kretzschmar, R., 2002. A survey of neural network classifiers for local wind prediction. Ph.D. Thesis, Institute for Signal and Information Processing, ETH Zurich.
- Kretzschmar, R., Eckert, P., Cattani, D., Eggimann, F., 2004. Neural network classifiers for local wind prediction. *Journal of Applied Meteorology* 43, 727–738.
- Landwehr, N.H.M., Frank, E., 2005. Logistic model trees. *Machine Learning* 59, 161–205.
- Le Cessie, S.v.H.J., 1992. Ridge estimators in logistic regression. *Journal of the Royal Statistical Society. Series C (Applied Statistics)* 41, 191–201.
- Lei, M., Shiyan, L., Chuanwen, J., Hongling, L., Yan, Z., 2009. A review on the forecasting of wind speed and generated power. *Renewable and Sustainable Energy Reviews* 13, 915–920.
- Lim, T., Loh, W., Shih, Y., 2000. A comparison of prediction accuracy, complexity, and training time of thirty-three old and new classification algorithms. *Machine Learning* 43, 203–228.
- Mohandes, M.A., Rehman, S., Halawani, O.T., 1998. A neural networks approach for wind speed prediction. *Renewable Energy* 13, 345–354.
- Quinlan, R., 1993. *C4.5: Programs for Machine Learning*. Morgan Kaufmann Publishers.
- Sallis, P.J., Shanmuganathan, S., Pavesi, L., Jarur, M., 2008. A system architecture for collaborative environmental modeling research. In: *Proceedings of the 2008 International Symposium on Collaborative Technologies and Systems (CTS 2008)*.
- Sfetsos, A., 2000. A comparison of various forecasting techniques applied to mean hourly wind speed time series. *Renewable Energy* 21, 23–35.
- Sumner, M., Frank, E., Hall, M., 2005. Speeding up logistic model tree induction. In: *Proceedings of the Knowledge Discovery in Databases Conference (PKDD 2005)*.