AUSTRIAN JOURNAL OF STATISTICS
Volume 31 (2002), Number 2&3, 169-175

Contribution to Skin Cancer Prevention in South Africa: Modelling the UV Index Utilizing Imprecise Data

Sep Human[1] and Vladimir B. Bajic[2]

[1]Centre for Systems Research,
Durban Institute of Technology, South Africa

[2]BioDiscovery Group,
Laboratories for Information Technology, Singapore

Abstract: South Africa has a high incidence of skin cancer and eye disorders because of the high number of sunshine hours per day. The ultraviolet (UV) index (UVI) provides a factual measure of the UV irradiance including biological effects. It is considered extremely important when gauging ultraviolet doses. A survey conducted across South Africa during January 1999 provided data records that contain nine independent variables for UVI inference purposes. This set of data includes cloud cover and other subjectively observed variables such as turbidity. The data set was recorded at 272 locations. Modelling the UVI by standard regression techniques using this data failed to produce reliable models for UVI prediction. The imprecision in some of the variables and small sample size implied that much more sophisticated techniques should be used. In the current research we resorted to artificial neural networks (ANNs) to cluster the data and then to model the UVI estimate in each of the data clusters. In the ANN training, we utilized the weights pruning method of optimal brain surgeon type to enable good generalization of the ANN models. The results obtained in the UVI assessment by this method produced results of significant accuracy.

Keywords: Skin Cancer, UV Index, Melanoma, Artificial Neural Networks, Imprecise Data, Modelling.

1 Problem Background

Exposure to sunlight and a fair skin are two major risk factors for skin cancer in humans. Risk levels for South Africans are amongst the highest worldwide. One in 45 white males, and one in 56 white females, had a lifetime risk of developing melanoma. By contrast, the lifetime risks in black males and females were one in 909 and one in 769, respectively. Taking all histologically diagnosed cancers in South Africa (SA), melanoma ranked 5th in white males (3.9%) and 3rd in white females (4.5%), 17th in black males (0.9%) and 13th in black females (1.2%). Rates for melanoma among South Africans are: white males $21.5/100\,000$, white females $17.8/100\,000$. Non-melanoma skin cancers show that whites ($196.8/100\,000$) had age adjusted rates comparable to those of Tasmania ($213.2/100\,000$), Sitas et al. (1998). Additionally, ambient ultraviolet (UV) radiation levels in SA are high throughout the year. The Southernmost town in SA, Cape Agulhas ($34°$S, $19°$E) is on about the same latitude as Perth and Sydney in Australia. UV indices greater than 10 are common, both in SA and Australia, Human and Bajic (1999). These extremely high

levels of UV radiation cause many health problems resulting in the high rates of skin cancer and eye disorders. These, on the other hand, could be considerably reduced, if a proper, country-wide information network on UV index (UVI) is provided.

In this article we present a simple, yet effective and cheap method, for assessing the UVI from imprecise data. Such data is simple to collect and also, collection process is simple to automate. The accuracy of UVI prediction in our experiments is very similar to that obtained by far more expensive systems used at national level, such as the one in Canada as presented in Long (1997).

2 Data

The purpose of the survey conducted in January 1999 was to record and assess a specific set of data (9 variables, including several estimated and categorical variables) and to infer the UVI from the collected data. Data was collected from 272 locations all over SA. These observation points were at popular beaches, urban, rural and almost featureless and very arid sites.

The following variables were collected: latitude, longitude, time, humidity, ambient temperature, cloud cover, albedo, turbidity and terrain. The last four were estimated and categorical. For example, a UVI of 15 was measured on the 28th of January at 12h 09min. The altitude was 1371 m (4500 ft) and the coordinates of Mnyokane (Swaziland) $26°$S, $31°$E. Cloud cover was 50% to less than 75%. The ambient temperature was $32.4°$ Celsius and the geography mountainous with reflective eroded cuttings (roadworks). Under a nearby tree the UVI had a range of 3 (max 4 and min 1).

3 Modelling

Preliminary data analyses by linear and nonlinear regression modelling did not produce satisfactory results for any reasonably accurate prediction of UVI. For example, with the allowed maximal prediction error of ± 4 units, linear regression model produced correct classification of only 31% of the data in the test set, while that score for logistic regression was 73%. Since the UVI values are highly nonlinearly dependent on the input data we have collected, the failure of the conventional regression modelling may be due to: 1/ a limited complexity of the resulting regression models, and 2/ the finite quantization ranges applied to imprecise data, such as albedo, cloud cover, terrain and turbidity, which are categorical variables and were subjectively assessed. To resolve the problem we used modelling based on artificial neural networks (ANNs) as they are known to be more flexible in modelling highly nonlinear relations among data, Bishop (1995), Zurada (1992).

3.1 Neural Network Model for UVI Assessment

The dataset contained 272 feature vectors of dimension $9,$ and 272 target values of the UVI. This set was randomly divided into the training set S_{tr} and the test set S_{tst}. The training set contained 174 input-target data, while the test set contained 98 input-target data. The data in S_{tr} were further split into 3 clusters using a self-organizing map (SOM)

Table 1: Content of the three data clusters

	$\mathbf{S}_{tr}$	$\mathbf{S}_{tst}$
# of data	66	34
# of data	35	24
# of data	73	40

ANN, Kohonen (1995). With such trained SOM network the set S_{tst} was also split into 3 clusters. Table 1 shows the number of input-target data in each of the clusters.

The following procedure is applied for each of the clusters. The training data was normalized to the zero mean and the standard deviation of 1. As the prediction model we selected a feedforward three-layered neural network (one hidden layer) containing $\tanh$ activation function of neurons $\left(\tanh(x) = \dfrac{e^x - e^{-x}}{e^x + e^{-x}}\right)$, Bishop (1995), in the hidden layer and a linear neuron in the output layer. This ANN was trained and pruned by the Optimal Brain Surgeon algorithm, Hassibi and Stork (1993), with regularization, Pederson et al. (1995), until the best performance has been achieved. The performance was estimated using the Akaike information criterion, Akaike (1973). Initially, networks had 60 neurons, and the number of neurons and the number of weights has been reduced through the pruning process. The network structure that had the highest value of the Akaike information criterion was then selected to represent the model for UVI assessment for data in that cluster. Finally, the trained networks were applied to the data in S_{tst}. By using the Mahalanobis distance, Bishop (1995), we observed that there is a small number of outliers in the datasets. When 18 most divergent outliers were removed from the datasets, the maximal prediction error was ± 3 units of the $0 - 15$ scale for the UVI. Note that 18 outliers represent only 6.62% of the whole dataset, so that the system was able to correctly classify 93.38% of all data with the maximal error of ± 3 units. With the elimination of the 38 outliers, which represent 13.97% of the whole original dataset, the maximal prediction error reduces to ± 2, while the correct classification is achieved for 86.03% of the original dataset. The distributions for the last case are represented in the bar graphs shown in Figs. 1-4. If elimination of such outliers is done, then an additional neural network can be trained on the whole cleaned dataset, so that the overall system can achieve very good prediction accuracy.

3.2 Comments

The results depict a significant prediction accuracy and precision of predictions. They are very comparable and actually better than that obtained with a much more complex and expensive system applied in Canada, Long (1997). Our model relies on a very simple technology for which data collection can be automated, so that a country-wide coverage by timely UVI predictions can be ensured. The two main reasons for not being able to further improve these results are, firstly, the available dataset is relatively small, and secondly, the imprecision present in some of the basic input data remains a problem for any technique used with the relatively small dataset we have. A more complete study, resulting in a much more complete data set, will reduce the modelling problem to a great

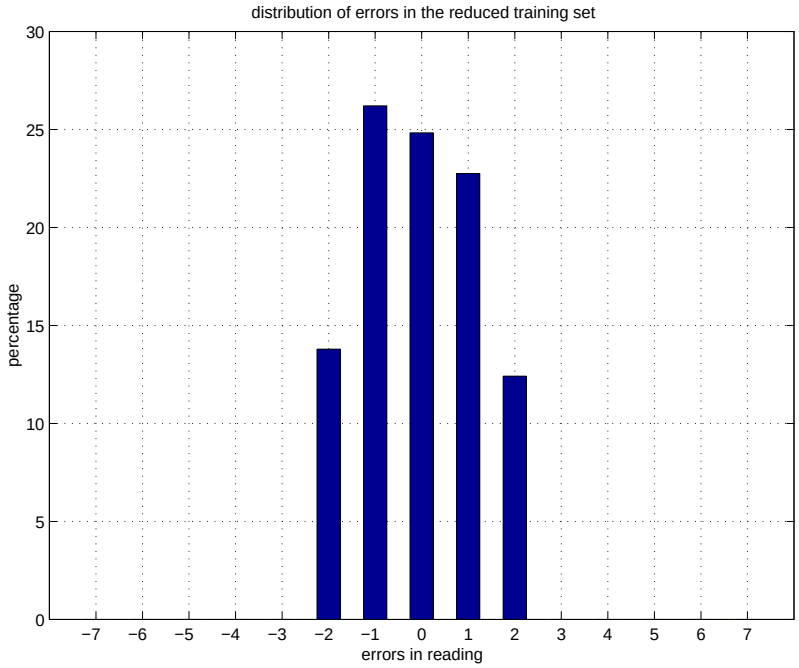

Figure 1: Distribution of errors in the training dataset with 38 outliers removed.

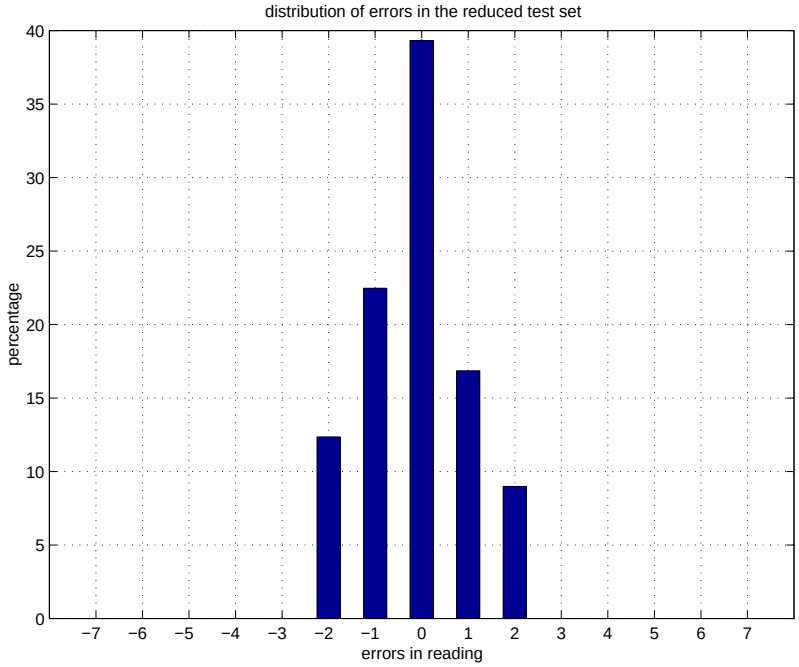

Figure 2: Distribution of errors in the test dataset with 38 outliers removed.

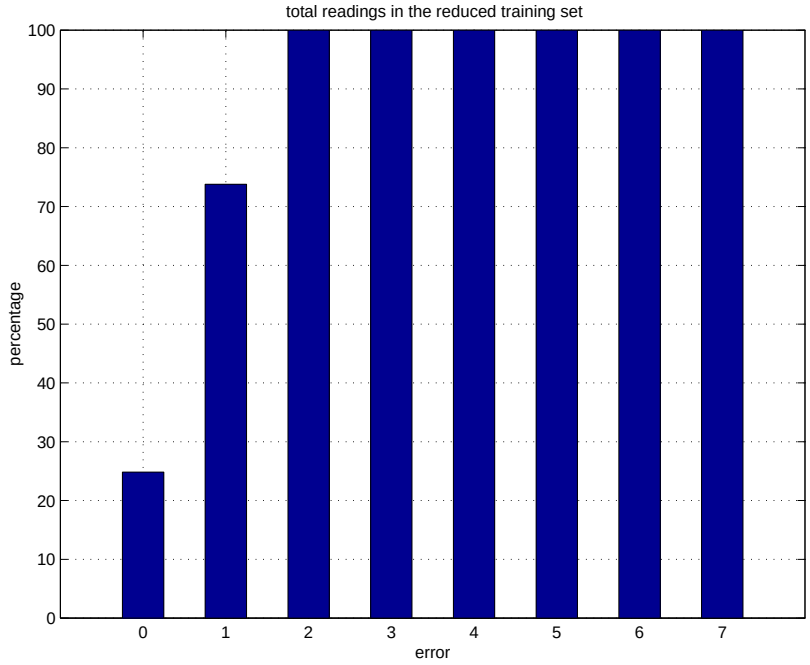

Figure 3: Cumulative distribution of errors in the training dataset with 38 outliers removed.

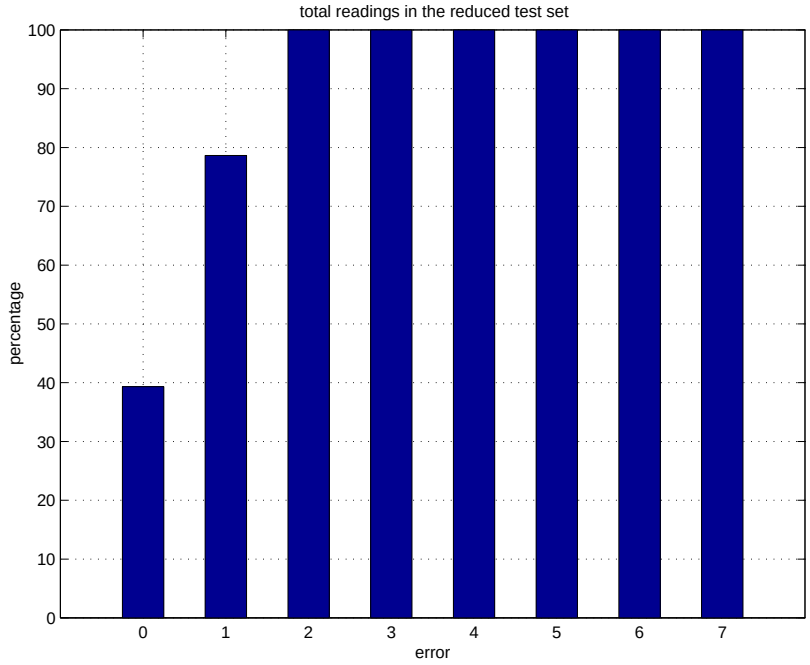

Figure 4: Cumulative distribution of errors in the test dataset with 38 outliers removed.

extent. Furthermore, a more sophisticated statistical and signal processing of the raw data can then be applied to further enhance the prediction results.

4 Conclusion

The paper presents an attempt to assess the UVI values from the imprecise data collected in South Africa. The results obtained are of a reasonably high accuracy, but can be improved by collecting more comprehensive dataset. Our study is directed at providing the assessment of UVI values based on easily collectable data so that these could be utilized country-wide as a guideline for population for adjusting sun-exposure behavior to minimize undesirable health effects stemming from UV overexposure.

References

H. Akaike. Information theory and an extension of the maximum likelihood principle. In B.N. Petrov and F. Csaki, editors, *Proc. 2nd International Symposium on Information Theory*, pages 267–281. Tsahkadsov, Armenia, USSR, 1973.

C.M. Bishop. *Neural Networks for Pattern Recognition*. Clarendon Press, Oxford, 1995.

B. Hassibi and D.G. Stork. Second order derivatives for network pruning: Optimal brain surgeon. In S.J. Hanson, J.D. Cowan, and C.L. Giles, editors, *Advances in Neural Information Processing Systems 5*, pages 164–171. Morgan Kaufmann, San Mateo, USA, 1993.

S. Human and V.B. Bajic. Estimation of uv index by neural networks. In V.B. Bajic, editor, *Development and Practice of Artificial Intelligence Techniques*, pages 75–77. IAAMSAD, South Africa, 1999.

T. Kohonen. *Self-Organizing Maps*. Springer-Verlag, Heidelberg, 1995.

J. Long. Monitoring uvi. In *Proc. of the Skin Cancer Conference*. Les Diablerets, Switzerland, 1997.

M.W. Pederson, L.K. Hansen, and J. Larsen. Pruning with generalization based weight saliences. In *Proc. of the Neural Information Processing Systems*, page 8. 1995.

F. Sitas, J. Madhoo, and J. Wessie. *Cancer in South Africa, 1993-1995*. 1998.

J.M. Zurada. *Introduction to Artificial Neural Systems*. West Publishing Company, St.Paul, Minnesota, 1992.

Authors' addresses:

Sep Human
Centre for Systems Research
Durban Institute of Technology
Box 953, 4000 Durban
South Africa

Tel. +27-82-955-0409
E-mail: SepH@ntech.ac.za

Professor Vladimir B. Bajic
BioDiscovery Group
Laboratories for Information Technology
21 Heng Mui Keng Terrace
Singapore 119613

Tel. +65-874-8800
E-mail: bajicv@lit.org.sg