Available online at www.sciencedirect.com

ScienceDirect

Transportation Research Procedia 3 (2014) 810 – 817

Transportation Research Procedia

www.elsevier.com/locate/procedia

17th Meeting of the EURO Working Group on Transportation, EWGT2014, 2-4 July 2014, Sevilla, Spain

A Neural Network based model for real estate price estimation considering environmental quality of property location

Vincenza Chiarazzo[a], Leonardo Caggiani[a]*, Mario Marinelli[a]

and Michele Ottomanelli[a]

[a]*Politecnico di Bari, via Orabona 4, Bari, 70125 Italy*

Abstract

In this paper, a model based on Artificial Neural Network (ANN) has been applied to real estate appraisal. Moreover, an evaluation of ANN performances in estimating the sale price of residential properties has been carried out. Artificial Neural Networks (ANNs) are useful in modelling input-output relationships learning directly from observed data. This capability can be very useful in complex systems like the real estate ones where motivations, tastes and budget availability often do not follow rational behaviours. This study also analyses the impact of such key environmental conditions that represent a problem related to many industrial cities where pollution and landscaping consequences affect the real estate market and residential location choices. We have considered a set of asking price's houses collected in the urban area of Taranto (Italy) where the biggest European steel factory and the 2nd industrial harbour are located.

Selection and peer-review under responsibility of the Scientific Committee of EWGT2014

Keywords: artificial neural networks; property appraisal; hedonic models; residential location choice.

1. Introduction

Recent theories on urban economy have highlighted how investments in transport system can improve accessibility to certain locations and affect property values. Moreover, city users are placing greater attention on the quality of the living environment. In polluted cities the environment quality has become an important attribute,

* Corresponding author. Tel.: +39-080-596-3380; fax: +39-080-596-3414.
E-mail address: leonardo.caggiani@poliba.it

Selection and peer-review under responsibility of the Scientific Committee of EWGT2014
doi:10.1016/j.trpro.2014.10.067

affecting residential and location choices and, consequently, land-use, mobility and economy of interested areas. This issue is important in industrial cities where people daily face the risk of diseases due to bad environmental conditions related to high pollution levels. In order to reduce this risk, people move to healthier residential areas which could be less accessible in terms of transportation system and/or urban services.

Hedonic studies related to real estate prices subject to transport conditions have completed urban economy theories and tested their hypotheses through different case studies (Ibeas et al., 2012, Chiarazzo et al., 2014).

The main aim of this work is to define a property appraisal model based on Artificial Neural Networks to estimate real estate prices. Moreover, some features affecting the real estate value have been determined paying particular attention to environmental factors such as pollution and noise levels, landscape, etc.

The proposed model could also help analysts in simulating interactions generated in an urban system where location choices for housing or companies strongly depend on the real estate market.

The main input parameters of the proposed model are transportation systems and environmental quality related attributes. In addition, input parameters widely used in the literature such as buildings characteristics and local land-use attributes have been also considered.

The paper is organized as follows. In section 2, the state of the art of the models developed to estimate the impacts on real estate is presented. In section 3, the proposed Artificial Neural Network (ANN) model is defined using a dataset from the urban area of Taranto (Italy). This study explores the impact of such key transport attributes and environmental elements including effects of a big steel factory and an industrial harbour which produce pollution and landscaping problems.

In section 4, the results are discussed. In the last section, some conclusions are carried out highlighting the opportunity introduced by ANN to capitalize environmental issues into housing market prices.

2. Literature review

Artificial Neural Networks (ANNs) are able to learn, to generalize results and to respond adequately to highly incomplete or previously unknown data (Shaw, 1992). ANN methodology was developed to capture functional forms, allowing the uncovering of hidden non-linear relationships between the variables. This method has been developed in the past years, especially using information of the study area showing outstanding performances. It represents a sub-field of computer science concerned with the use of computers in tasks that are normally considered to require knowledge and cognitive abilities (Gevarter, 1985). It has been applied to the property price forecasting in recent years (Lai Pi-ying, 2011).

Borst (1991) has defined a great number of variables in his network to appraise real estate in New York State, demonstrating that ANNs are able to predict the real estate price with 90% accuracy.

ANNs perform better than multi-variate analysis, since networks are nonlinear. They can also evaluate subjective information, such as the transport system and the characteristics of the zone, which are difficult to incorporate into traditional mathematical approaches.

Traditional multiple regression models have focused on the relationship between real estate prices and accessibility until a systematic review of various research works was carried out by Fujita (1989). Research about how transport system can influence real estate prices was initiated by von Thünen (1826) who laid the foundations of a theory about the distribution of land use and rents in urban areas as proposed by Alonso (1964), Muth (1969) and Mills (1972).

Many hedonic studies have specified the role of quality of environment considering the real estate price (Din et al. 2001), accessibility and other local land-use attributes (Ibeas et al., 2012; Chiarazzo et al., 2014).

The results highlight a significant influence of variables such as the distance in kilometres to reach the industrial centre or the value of environmental polluters on the variability of the relationship between accessibility to bus stop and real estate prices. The properties located close to the industrial area also showed significant and negative changes in value (Chiarazzo et al., 2014).

Other studies have focused on the impact resulted by Bus Rapid Transit systems on real estate prices (Rodríguez and Mojica, 2009). These studies showed the impact on property values resulted by introducing a Bus Rapid Transit (BRT) system in a city and found an increase in price.

In this paper, an ANN approach is proposed with an analysis of performances in estimating the sale price of residential properties.

3. Artificial Neural Network models and the data set

Neural networks can be used to predict the sale price of a house. From this point of view, the core of this paper relies on the application of Artificial Neural Networks to real estate appraisal. An advantage of the proposed approach is that, using ANNs, there is no need to assume explicit functions between input and output of the studies because an ANN learns directly from observed data.

The ANN model used in this study has been trained with data gathered from the city of Taranto (Italy), which represents a high environmental risk area, as declared by the Italian Ministry of Environment since 1991.

In particular, Taranto urban area (see Fig. 1) is characterized by a location migration flow towards zones far from the city centre which are close to the industrial plant and the harbour.

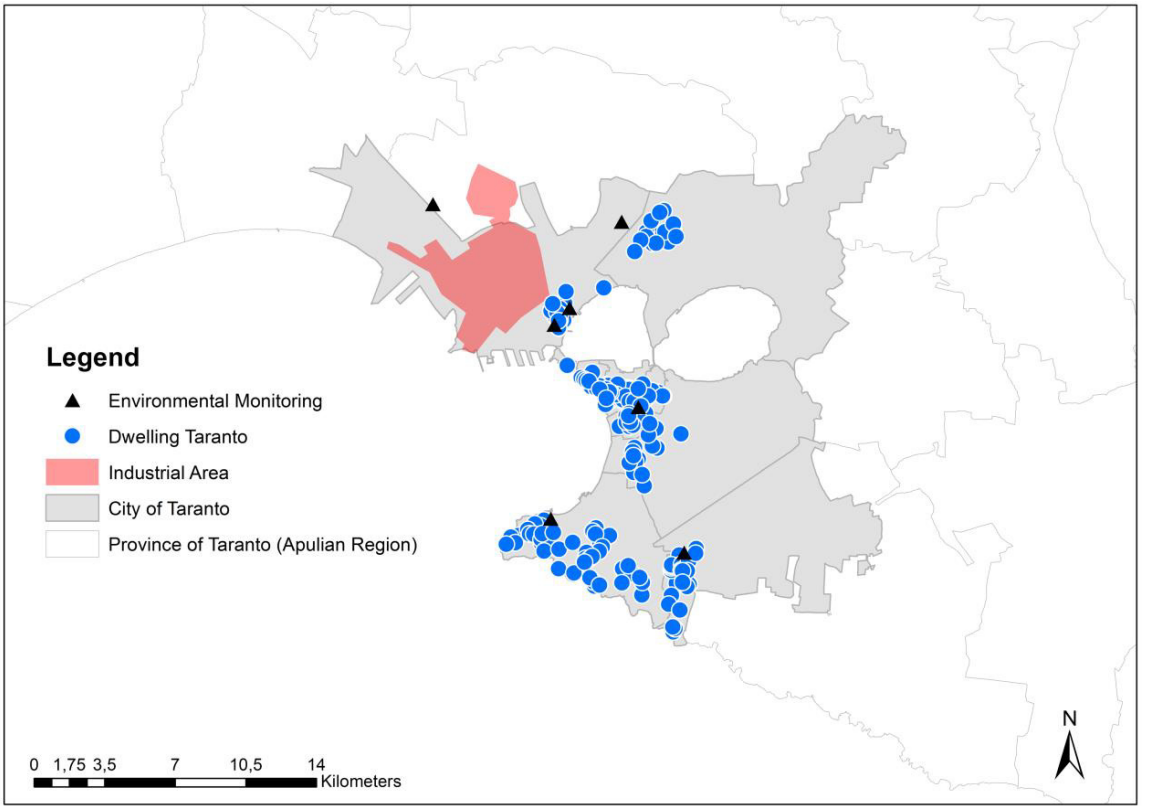

Fig. 1. Location of sampled dwellings, industrial area and environmental monitoring in the study area.

Furthermore, as resulting from the Local Transport master plan, chosen locations are characterized by low accessibility and services without relevant benefit in terms of real estate market prices.

The household sample comes from a cross sectional database obtained through various on-line real estate platforms (October 2012). Data on environmental pollution levels have been measured and provided by the Regional Agency for the Prevention and Protection of the Environment of Apulia Region (ARPA) together with data from the urban habitat database of a Geographic Information System (GIS).

Starting from the variables reported into the residential property database, the ANN has been trained in order to relate environmental and other considered local attributes to the property prices. The considered ANN is a feed-forward back-propagation network (Hagan and Menhaj, 1994) with a training function that updates weight and bias values according to Levenberg-Marquardt optimization.

The ANN is composed by three hidden layers containing 20 neurons in the first and the second one and one neuron in the last one. The input data to train the ANN consist of matrices containing environment and land use attributes of the collected properties. The output is represented by the real estate price.

The dataset (with 193 records) has been divided into three different subsets: a training set (70% of total records) a validation set (15%) and a test set (15%). These three subsets have been built in order to include in each of them all the price ranges of the considered properties.

The following variables were collected for each property and coded into the database (see Table 1):

- PAP is the property asking price;
- SQM is the surface area of the property in square meters;
- ROOMS is the number of bedrooms at the property;
- BATH is the number of bathrooms at the property;
- IMPR is a dummy variable taking a value of 1 if the property requires major improvement;
- FLOOR is the floor where the property is located in the building;
- LIFT is a dummy variable taking a value of 1 if the building where the property is located has a lift (elevator);
- TAP is a dummy variable taking a value of 1 if the property is a flat;
- TSAP is a dummy variable taking a value of 1 if the property is a single family flat;
- TRH is a dummy variable taking a value of 1 if the property is a rural house;
- TTF is a dummy variable taking a value of 1 if the property is a two-storey flat;
- TSF is a dummy variable taking a value of 1 if the property is a studio flat;
- TE is a dummy variable taking a value of 1 if the property has a terrace;
- ILVA is the distance in kilometres to reach the industrial center from the property using the road network;
- GA is a dummy variable taking a value of 1 if the property has a garden;
- BCH is a dummy variable taking a value of 1 if the property is located at a beach at less than 800 m from a beach;
- GAR is a dummy variable taking a value of 1 if the property has a garage;
- NC is a dummy variable taking a value of 1 if the property is a new construction;
- LTI is the number of internal lines bus serving the zone;
- FTV is a dummy variable taking a value of 1 if the property has a bus stop less than 400 m away interacting with the number of lines servicing that bus stop;
- TACCT is the time in minutes which it takes at morning rush hour to reach Taranto's CBD from the property using the road network, considering congestion;
- TRAIN is a dummy variable taking a value of 1 if the property is less than 500 m from a suburban train station;
- CE is a dummy variable taking a value of 1 if the property is located in the city center;
- DEN is a measure of the population density in a zone (inhabitants per area unit);
- HOU is the number of zone households;
- ENT is the number of zone enterprises;
- EMP is the number of zone employees;
- EMPRES is the number of employed residents in a zone;
- INH is the number of inhabitants in a zone;
- INHAGG is the number of inhabitants in an aggregate zone;
- AREA is the zone area;
- SO2, NOX, NO, NO2, CO, PM10 are values of measured environmental polluters (μg/m3 293K);
- SO2_MAX, NOX_MAX, NO_MAX, NO2_MAX, CO_MAX, PM10_MAX are maximum values of measured environmental polluters (μg/m3 293°K). The pollutants are measured by eleven fixed air quality stations located in the study area giving out also maximum values of the pollutants.

Table 1. Variables description.

Variable	Minimum	Maximum	Mean	Std. deviation	Measurement unit
PAP	39000	370000	140193	77644.11	*EUR*
SQM	32	1004	116.64	85.14	m^2
ROOMS	1	9	3.45	1.40	*No. of rooms*
BATH	0	3	1.48	0.58	*No. of bathrooms*
IMPROV	0	1	0.24	0.43	-
FLOOR	1	3	1.14	0.37	*Floor number*
LIFT	0	1	0.31	0.46	*Elevator*
TAP	0	1	0.70	0.46	-
TSAP	0	1	0.19	0.39	-
TRH	0	1	0.02	0.13	-
TTF	0	1	0.05	0.23	-
TSF	0	1	0.04	0.19	-
TE	0	1	0.11	0.31	-
ILVA	1	31.80	8.38	4.78	*km*
GA	0	1	0.40	0.49	-
BCH	0	1	0.17	0.38	-
GAR	0	3	0.32	0.47	-
NC	0	1	0.04	0.19	-
LTI	3	5	3.83	0.93	*No. of internal lines*
FTV	1	7	2.21	1.19	*Minutes*
TACCT	1	23	11.01	6.43	*Minutes*
TRAIN	0	1	0.45	0.50	-
CE	0	1	0.37	0.49	-
DEN	0.01	4.18	0.51	1.07	*Inhabitants/*m^2
HOU	0	367	120.07	94.82	*No. of households*
ENT	0	58	11.70	12.33	*No. of enterprises*
EMP	0	2980	61.77	253.31	*No. of employees*
EMPRES	0	394	85.14	83.08	*No. of employed*
INH	0	1255	354.42	293.06	*No. of inhabitants*
INHAGG	15849	31214	24657.80	4525.38	*No. of aggr. inhabitants*
AREA	6636.99	1884885.55	463247.32	346001.77	m^2
SO2	0.54	3.54	1.58	0.84	$\mu g/m^3$ *293K*
NOX	12.37	72.83	40.87	25.13	$\mu g/m^3$ *293K*
NO	2.20	21.97	11.33	7.41	$\mu g/m^3$ *293K*
NO2	9.01	49.40	25.74	16.48	$\mu g/m^3$ *293K*
CO	0.29	0.54	0.37	0.12	$\mu g/m^3$ *293K*
PM10	28.36	35.85	34.75	2.46	$\mu g/m^3$ *293K*
SO2_MAX	10.83	141.80	43.00	38.39	*Maximum value of* SO_2
NOX_MAX	166.49	1102.42	484.21	320.93	*Maximum value of* NO_x
NO_MAX	102.54	638.19	312.43	220.84	*Maximum value of NO*
NO2_MAX	74.86	254.95	149.47	70.95	*Maximum value of* NO_2
CO_MAX	3.72	5.82	4.43	1.00	*Maximum value of CO*
PM10_MAX	101.66	199.27	114.09	29.49	*Maximum value of PM10*

4. Results

At the end of the training procedure, for each input sample, an output (estimated house price) has been estimated by the ANN model. Then, the obtained results have been compared with the target values (actual house price).

The correlation coefficient (R-value) between the outputs and targets values has been chosen as a fit index to evaluate the goodness of the trained ANN. R-values have been calculated for the training set (Train), the validation set (Valid), the test set (Test) and the entire dataset (All).

For the proposed ANN model, the R-value related to the training set is very close to 1, which indicates a very good fit. The R-value related to the test set is 0.83, which indicates a good fit, whereas the data used did not belong to the dataset used for the training phase.

Figure 2 illustrates the obtained output, target and R-values. In this figure the ANN outputs are plotted versus the targets (circles). The best linear fit is indicated by the solid line. The perfect fit (output equal to targets) is indicated by a dashed line.

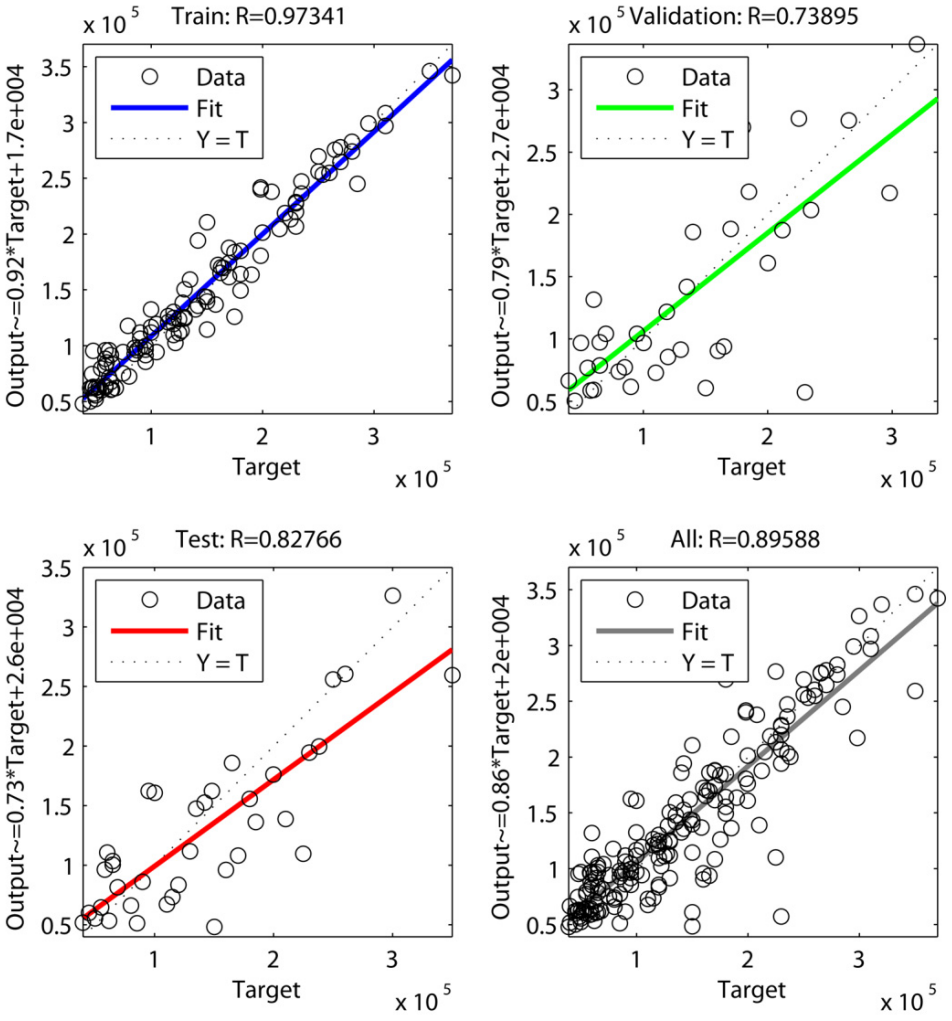

Fig. 2. Performances of the ANN model.

In order to evaluate the most significant input variables, a sensitivity analysis has been carried out. For this purpose, starting from the same dataset, the ANN training phase has been repeated 42 times, eliminating each time one of the 42 input variables. The significance of each removed input variable was evaluated in function of the R-value obtained at the end of each training procedure.

In other words, as the R-value decreases, the significance of the removed variable increases. Analysis results are reported in Table 2 where variables are listed in descending order of significance.

Table 2. Sensitivity analysis results.

Significance order	Variable	R-value (All)	Significance order	Variable	R-value (All)	Significance order	Variable	R-value (All)
1	BCH	0.001	15	LTI	0.800	29	LIFT	0.839
2	GA	0.440	16	TRAIN	0.802	30	TRH	0.839
3	TE	0.505	17	FTV	0.802	31	NOX_MAX	0.840
4	EMP	0.584	18	EMP	0.804	32	TSAP	0.846
5	BATH	0.657	19	ROOMS	0.812	33	AREA	0.847
6	TAP	0.668	20	TTF	0.813	34	NO	0.848
7	ILVA	0.716	21	CO_MAX	0.820	35	PM10	0.851
8	SO2_MAX	0.752	22	NO2_MAX	0.823	36	HOU	0.852
9	NC	0.756	23	CO	0.824	37	ENT	0.856
10	GAR	0.777	24	PM10_MAX	0.825	38	TACCT	0.856
11	DEN	0.786	25	NO2	0.828	39	INH	0.864
12	CE	0.789	26	SO2	0.832	40	NOX	0.867
13	FLOOR	0.792	27	SQM	0.833	41	EMPRES	0.873
14	NO_MAX	0.798	28	TSF	0.838	42	INHAGG	0.887

The most significant variables appear to be those related to property special features such as proximity to the beach or presence of a garden or terrace. Moreover, the most significant environmental variable is the maximum value of SO_2. Transportation variables (LTI and TRAIN) are also quite relevant. On the other hand, the variable that does not affect the model is the number of inhabitants in aggregate zones (INHAGG) because its absence does not change the R-value of the model. We have to point out that the variable ILVA, which describes the distance between the industrial center and the property location, is ranked among the most significant ones. For this reason we have also defined an ANN trained without the group of environmental polluters variables. As a result, we obtained an R-value equal to 0.819. This value is lower than that achieved with all the involved 42 variables, but it is still good since it arises from a simpler model. This result shows that the distance between the industrial center and a property location (more easily perceived by the population if compared to the environmental polluters values) is well suited as a synthetic parameter of environmental quality.

5. Conclusions

Artificial Neural networks can be used to predict the house sale price. This work highlights the role of the environment quality considering the real estate price, accessibility and other local land-use variables. The proposed model has been calibrated using data collected in the city of Taranto. Furthermore, a sensitivity analysis has been carried out in order to identify the most significant input variables. The traditional multiple regression models and the estimated neural network models were useful in highlighting how different transport characteristics as well as the environmental quality affect the prices of real estate properties. The proposed model has a good fit in both training and test results. The results provided by the neural network can support investments in transport system. The ANN model can help appraisers in making assessments and environmental regeneration. However, further investigations are ongoing. In particular, clustering methods will be applied to improve the statistical performance of the ANN in order to capture specific characteristic of groups of properties.

Acknowledgements

Authors wish to thank the Regional Agency for the Prevention and Protection of the Environment of Apulia Region (ARPA - Agenzia Regionale per la Prevenzione e la Protezione dell'Ambiente) for providing the data on environmental pollution in the study area.

References

Alonso, W., 1964. Location and Land Use: Toward a General Theory of Land Rent. Harvard University Press, Cambridge.

Borst, R.A., 1991. Artificial Neural Networks: The Next Modeling/Calibration Technology for the Assessment Community? Property Tax Journal (International Association of Assessing Officers), 10(1), 69–94.

Chiarazzo, V., Dell'Olio, L., Ibeas, A., Ottomanelli, M., 2014. Modeling the effects of environmental impacts and accessibility on real estate prices in industrial cities. Procedia - Social and Behavioral Sciences, 111, 460–469.

Din, A., Hoesli, M., Bender, A., 2001. Environmental variables and real estate prices. Urban Studies, 38(11), 1989–2000.

Fujita, M., 1989. Urban Economic Theory: Land Use and City Size. Cambridge University Press, Cambridge.

Hagan, M.T., M. Menhaj, 1994. Training feed-forward networks with the Marquardt algorithm. IEEE Transactions on Neural Networks, 5(6), 989–993.

Ibeas, T., Cordera, R., dell'Olio, L., Coppola, P., Dominguez, A., 2012. Modelling transport and real-estate values interactions in urban systems. Journal of Transport Geography, 24, 370–382.

Lai Pi-ying, 2011. Analysis of the Mass Appraisal Model by Using Artificial Neural Network in Kaohsiung City. Journal of Modern Accounting and Auditing, 7(10), 1080–1089.

Shaw, J., 1992 . Neural network resource guide. AI Expert, 8(2), 48–54.

Gevarter, W. B., 1985. Intelligent machines: An introductory perspective of artificial intelligence and robotics. Prentice Hall Inc, New Jersey.

Mills, E.S., 1972. Studies in the Structure of the Urban Economy. Johns Hopkins Press, Baltimore.

Muth, R.F., 1969. Cities and Housing: The Spatial Pattern of Urban Residential Land Use. University of Chicago Press, Chicago.

Rodríguez, D.A., Mojica, C.H., 2009. Capitalization of BRT network expansions effects into prices of non-expansion areas. Transportation Research Part A, 43, 560–571.

von Thünen, J.H., 1966. Der isolierte staat in beziehung auf landwirtschaft und nationalökonomie. Oxford University Press, Oxford.