RULE-BASED AND CASE-BASED REASONING IN HOUSING PRICES

By

Gabrielle Gayer, Itzhak Gilboa and Offer Lieberman

November 2004

COWLES FOUNDATION DISCUSSION PAPER NO. 1493

COWLES FOUNDATION FOR RESEARCH IN ECONOMICS

YALE UNIVERSITY

Box 208281

New Haven, Connecticut 06520-8281

http://cowles.econ.yale.edu/

Rule-Based and Case-Based Reasoning in Housing Prices*

Gabrielle Gayer†, Itzhak Gilboa‡, and Offer Lieberman§

October 2004

Abstract

People reason about real-estate prices both in terms of general rules and in terms of analogies to similar cases. We propose to empirically test which mode of reasoning fits the data better. To this end, we develop the statistical techniques required for the estimation of the case-based model. It is hypothesized that case-based reasoning will have relatively more explanatory power in databases of rental apartments, whereas rule-based reasoning will have a relative advantage in sales data. We motivate this hypothesis on theoretical grounds, and find empirical support for it by comparing the two statistical techniques (rule-based and case-based) on two databases (rentals and sales).

1 Introduction

1.1 Motivation and Hypothesis

How do people assess real estate prices? Casual observation suggests that two modes of reasoning are very common in generating such assessments. The

*We wish to thank Don Brown for the conversations that greatly influenced this work. We thank the Student Association of Tel-Aviv University and Professor Juval Portugali for the data. This work was supported by a grant from the Sapir Center at Tel-Aviv University.

†Tel-Aviv University. gajer@post.tau.ac.il

‡Tel-Aviv University and Yale University. igilboa@post.tau.ac.il. Gilboa gratefully acknowledges ISF grant no. 975/03.

§The Technion, Israel Institute of Technology. offerl@ie.technion.ac.il

first relies on general rules, such as, "In this area, the price per squared meter is \$3,000". The second is case-based, as in the argument, "The apartment next door, practically identical to mine, was just sold for \$300,000". Indeed, in the US the standard assessment procedure involves two assessments, one that is rule-based and another that is case-based.

It seems safe to assume that, for the most part, both types of reasoning are present when a person attempts to assess the market price of a given real-estate asset. The question we wish to address is whether one can make any qualitative predictions regarding the relative importance of rule-based versus case-based reasoning. Specifically, do people think differently about apartments for sale and apartments for rent?

We hypothesized that the answer would be in the affirmative. The reason is as follows. A rental apartment is a pure consumption good. When one is asked to assess the market price of such an apartment one may be using both case-based and rule-based reasoning. Let us take this mix as a benchmark, and ask how would the reasoning change if the apartment were for sale.

An apartment for sale is partly a consumption good, and partly an investment. Its value, should one wish to re-sell it, is determined by the market. It follows that a person who considers buying an apartment needs to worry not only about how much the apartment is worth to her, but also how much it is worth to others. The purchase of apartment becomes a coordination game of sorts: to a large extent, an apartment is worth what people think it is worth, namely, whatever price the market coordinates on. Assessing the rent of an apartment does not have this coordination aspect, unless one intends to sublet the apartment.

We maintain that rule-based pricing is easier for the market to coordinate on than is case-based pricing. The reason is that rules are simple to state and to transmit, whereas cases are numerous and difficult to convey. To illustrate this point, imagine that an experienced real-estate agent wishes to transfer her knowledge to a young colleague. If this knowledge takes the

form of a rule, it will generally be succinct and easily stated. If, however, the expert's knowledge is case-based, it is necessary to convey the expert's similarity function, but also the entire database of cases that she uses for generating assessments. It follows that rules, which are by nature succinct and easy to describe, are easier to coordinate on than are cases. We therefore hypothesize that case-based reasoning will have a relative advantage in explaining rental data, whereas rule-based reasoning will have a relative advantage in explaining sales data.

1.2 Methodology

We analyze two databases of asking prices on apartments in the greater Tel-Aviv area: one consists of apartments for rent, and the other – for sale. We contrast the simplest possible models of rule-based and of case-based reasoning. Rule-based reasoning is represented by hedonic regression (see Rosen (1974)), where the asking price is regressed linearly on certain characteristics of the apartment such as its size, number of rooms, floor, etc. If we denote the asking price in observation i by Y_i and the vector of characteristics – by $X_i = (X_i^1, ..., X_i^m)$, we estimate the regression

$$Y_i = \beta_0 + \beta_1 X_i^1 + ... + \beta_m X_i^m + \varepsilon_i \tag{1}$$

How should we model case-based assessments, and how should we estimate such a model? Gilboa, Lieberman, and Schmeidler (2004) axiomatize an assessment rule that is based on a similarity function $s : \mathbb{R}^m \times \mathbb{R}^m \to \mathbb{R}_{++}$. Given such a function s, n observations $(X_i^1, ..., X_i^m, Y_i)$ for $i = 1, ..., n$, and a new apartment with characteristics $X_{n+1} = (X_{n+1}^1, ..., X_{n+1}^m)$, they suggest that Y_{n+1} be assessed by the similarity-weighted average of past Y_i values. More explicitly,

$$Y_{n+1} = \frac{\sum_{i \le n} s(X_i, X_{n+1}) Y_i}{\sum_{i \le n} s(X_i, X_{n+1})} + \varepsilon_{n+1} \tag{2}$$

where $\varepsilon_i \overset{i.i.d.}{\sim} N(0, \sigma^2)$.

This formula should be interpreted as follows. Ms. A wants to sell her apartment, with characteristics $X_{n+1} = (X^1_{n+1}, ..., X^m_{n+1})$. She has to determine her asking price, Y_{n+1}. She gets to observe the asking prices on other, similar apartments, Y_i, $i = 1, ..., n$. She evaluates the similarity between the characteristics of her apartment, X_{n+1}, and the characteristics of each apartment she has seen on the market, X_i. This similarity is $s(X_i, X_{n+1})$. Next, Ms. A decides that a reasonable asking price for her apartment will be the similarity-weighted average of the asking prices she has observed, where the price Y_i gets a weight proportional to the similarity of apartment i to apartment $n + 1$. As usual, the error term ε_{n+1} stands for various unobservable variables, inherent uncertainty, and measurement errors.

Suppose that equation (2) models the way people determine asking prices. We would now like to estimate the function s from the data, in a way that parallels the estimation of the coefficients $(\beta_j)_{0\leq j\leq m}$ in linear regression. To this end, we would like to assume that an equation such as (2) governed the process that generated $(Y_t)_{t\leq n}$. However, the data we have are not ordered. Therefore, in the estimation process we assume that each Y_t is distributed around the weighted average of all other values, $(Y_i)_{i\neq t}$. Specifically,

$$Y_t = \frac{\sum_{i\neq t} s(X_i, X_t)Y_i}{\sum_{i\neq t} s(X_i, X_t)} + \varepsilon_t \qquad \text{for every } t \leq n \tag{3}$$

Observe that we assume that the function s is the same for all individuals who generated past data $(Y_t)_{t\leq n}$. This assumption parallels the assumption in equation (1), that the coefficients $(\beta_j)_{0\leq j\leq m}$ are independent of i.[1]

Estimating the function s from a given database is consistent with a scenario in which all sellers have access to exactly the same database, which is also the one we analyze. This would be the case if all sellers obtained the same database that we have, and, more importantly, had no access to asking prices of other sellers posted in other databases. This assumption is,

[1] Alternatively, one may view our approach as estimating a similarity function of a representative agent, as axiomatized in Gilboa, Lieberman, and Schmeidler (2004).

of course, not very realistic. Moreover, in reality we cannot expect to have access to the actual database that each and every seller has. Hence, we take the single database that we have as a proxy for the databases that each seller had. Should our database be representative of the information that sellers actually have, we might hope that the estimation process will be unbiased.

The equations (3) do not suffice to specify the values of $(Y_t)_{t\leq n}$ as a function of $(\varepsilon_t)_{t\leq n}$. These equations can be solved to extract the differences between any two Y_t's. But if $(Y_t)_{t\leq n}$ solve (3), so would $(Y_t+\lambda)_{t\leq n}$ for every $\lambda\in\mathbb{R}$. We therefore add a parameter α to the model, which will stand for the expected value of $(Y_t)_{t\leq n}$. The resulting model is:

$$\sqrt{n}\left(\bar{Y}_n-\alpha\right)=\varepsilon_1$$

where

$$\bar{Y}_n=\frac{1}{n}\sum_{i\leq n}Y_i$$

and, for every $1<t\leq n$,

$$Y_t=\frac{\sum_{i\neq t}s(X_i,X_t)Y_i}{\sum_{i\neq t}s(X_i,X_t)}+\varepsilon_t \tag{4}$$

In this paper we take a parametric approach to the estimation of the function s in the system (4). The advantages of a parametric approach in our case are threefold. First, a parametric approach simplifies the analysis. Second, it serves as a reasonable counterpart to the parametric approach of linear regression, and allows a comparison of two models with the same number of unknown parameters. Finally, our parametric approach will also allow us to test hypotheses about the significance of particular variables in the similarity model (4), in a way that parallels the tests of significance in the regression model (1).

Specifically, we are interested in similarity functions that depend on a weighted Euclidean distance. Define, for a vector $w\in\mathbb{R}^m_{++}$, the w-weighted

squared Euclidean distance:

$$d_w(x, x') = \sum_{j \leq m} w_j (x_j - x'_j)^2 \tag{5}$$

This function allows different variables to have different impact on the measure of "distance". There are two reasons for which we resort to a weighted Euclidean distance rather than, say, standard Euclidean distance. First, the variables are on different scales. For instance a difference of 1 in "number of rooms" is quite different from a difference of 1 in "area in square feet". Second, even if the variables were normalized, a variable such as "number of rooms" would probably be more influential than a variable such as "the apartment has bars on its windows". The weighted Euclidean distance allows a wide range of distance functions, weighing the relative importance of the variables involved.

Next, we wish to translate the distance function to a similarity function. It is natural to assume that the similarity function is decreasing in the distance, and as the distance goes up from 0 to ∞, the similarity function goes down from 1 (maximal similarity) to 0. We define the similarity function by

$$s_w(x, x') = \frac{1}{1 + d_w(x, x')} \tag{6}$$

Plugging this function into the system (4) we obtain the parametric version of our model, which we estimate. We will henceforth refer to (4) with the additional specification $s = s_w$.

Given estimators $(\hat{\beta}_j)_{0 \leq j \leq m}$ of the parameters $(\beta_j)_{0 \leq j \leq m}$ in equation (1), and estimators $(\hat{w}_j)_{1 \leq j \leq m}$ of the parameters $(w_j)_{1 \leq j \leq m}$ in equation (4), we can ask which model fits the data better, for each of the databases we analyze. Observe that the two models have exactly the same number of parameters, namely, $m+2$ (including σ^2). We wish to compare the two models in terms of their likelihood functions, as well as in terms of the out-of-sample predictions generated by their maximum likelihood estimators. To this end we need

to compute the likelihood function of (4). Maximization of this likelihood function will provide an estimate of the weights $(w_j)_{1\leq j\leq m}$, and will also allow us to test them for significance, in a way that parallels significance tests for $(\beta_j)_{0\leq j\leq m}$ in linear regression.

1.3 Related Literature

Hedonic regression has been a standard tool for studying real-estate pricing for decades (see Rosen, 1974). Spatial methods have also been well-established and widely used tools. (See Ord, 1975, Ripley 1981, 1988, Anselin, 1988, and Dubin, 1988.) A typical model would regress the price variable on several hedonic variables, as well as on other price variables, in a manner that bears mathematical resemblance to autocorrelation techniques. Specifically, whereas in an autocorrelation model a variable Y_t is regressed on its past values $Y_{t-1}, Y_{t-2}, ...$, in spatial models real-estate properties that are geographically close are assumed to be interrelated. Recent models of this type include Kim, Phipps, and Anselin (2003) and Brasington and Hite (2004), who use the following model

$$\nu = \rho W\nu + X\beta + W\underline{X}\alpha + \varepsilon, \qquad \varepsilon \sim N\left(0, \sigma^2 I_n\right) \tag{7}$$

where W is a fixed, known matrix.[2] Thus, in this model, the price vector ν depends on a weighted sum of itself, $W\nu$. In this respect, our model (4) resembles (7). However, in (7), the matrix W is assumed fixed, whereas we derive it from a similarity function and estimate this function.

The regression model we use in this paper is a classical example of the hedonic regression family. It is much simpler than spatial regression models such as (7). By contrast, our similarity model does not seem to have a counterpart in the literature. It differs from spatial regression models in two

[2] The efficacy of purely hedonic and of spatial regression models has also been a topic of study. (See Gao, Asami, and Chung, 2002.)

important ways. First, as mentioned above, in our model the similarity matrix is estimated empirically. Second, our model uses a similarity-weighted average formula, rather than a linear formula, as the underlying data generating process.

Our goal in this paper is to compare two modes of reasoning, represented by two statistical methodologies. To this end, we chose to use exactly the same variables and the same number of parameters in each model. In doing so, we also provided each model with equally low levels of preliminary reasoning or external information. It stands to reason that one may combine the two models in a way that parallels the combination of rule-based and case-based techniques in human reasoning, and thereby to obtain a better fit for the data than either model can achieve on its own.

The paper is organized as follows. Section 2 develops the statistical theory. It computes the likelihood function for the model (4), and develops tests for the significance of weights $(w_j)_{1\leq j\leq m}$. Section 3 describes the data, the analysis conducted, and the results. It also comments on some statistical issues that arise in the interpretation of the results. Section 4 concludes with final remarks.

2 Statistical Theory

2.1 The Likelihood function

Define

$$S = S(w) = \begin{pmatrix} 1/\sqrt{n} & 1/\sqrt{n} & ... & 1/\sqrt{n} \\ -\frac{s_{w,2,1}}{\sum_{i\neq 2} s_{w,2,i}} & 1 & & -\frac{s_{w,2,n}}{\sum_{i\neq 2} s_{w,2,i}} \\ ... & & ... & \\ -\frac{s_{w,n,1}}{\sum_{i\neq n} s_{w,n,i}} & & -\frac{s_{w,n,n-1}}{\sum_{i\neq n} s_{w,n,i}} & 1 \end{pmatrix}.$$

The structural and reduced form models are

$$Sy = \sqrt{n}\alpha e_1 + \varepsilon$$

and

$$y = \sqrt{n}\alpha S^{-1}e_1 + S^{-1}\varepsilon.$$

where e_i is the i-th unit vector, y and ε are $n \times 1$ vectors, with $\varepsilon \sim N(0, \sigma^2 I)$. Note that $S1 = \sqrt{n}e_1$, where 1 is the $n \times 1$ vector whose entries are all 1. Hence $S^{-1}e_1 = n^{-1/2}1$, so that

$$y = \alpha 1 + S^{-1}\varepsilon.$$

That is, the unconditional expectation of the y-vector is α.

Set

$$H = \frac{S'S}{\sigma^2}.$$

The log-likelihood function is

$$l(\theta) = -\frac{n}{2}\log(2\pi) + \frac{1}{2}\log\det(H) - \frac{1}{2}(y - \alpha 1)'H(y - \alpha 1).$$

Clearly, for any given $(w_j)_{1 \le j \le m}$, the profile MLE of α is

$$\hat{\alpha} = (1'H1)^{-1}1'Hy = \bar{Y}_n,$$

since $1'S' = \sqrt{n}e_1'$.

Define

$$S_0 = S_0(w) = \begin{pmatrix} 0 & 0 & \cdots & 0 \\ -\frac{s_{w,2,1}}{\sum_{i \ne 2} s_{w,2,i}} & 1 & & -\frac{s_{w,2,n}}{\sum_{i \ne 2} s_{w,2,i}} \\ \cdots & & \cdots & \\ -\frac{s_{w,n,1}}{\sum_{i \ne n} s_{w,n,i}} & \cdots & -\frac{s_{w,n,n-1}}{\sum_{i \ne n} s_{w,n,i}} & 1 \end{pmatrix}.$$

Now, $Sy - \sqrt{n}\bar{Y}_n e_1 = S_0 y$. The profile log-likelihood function is readily seen to be

$$l_P(w) = -\frac{n}{2}\left[\log(2\pi) + 1 - \log n\right] - \frac{n}{2}\log\left(y'S_0'(w)S_0(w)y\right) + \frac{1}{2}\log\det\left(S'(w)S(w)\right).$$

It follows that the log-likelihood function will be maximized for $(w_j)_{1 \le j \le m}$ that maximize

$$-\frac{n}{2}\log\left(y'S_0'(w)S_0(w)y\right) + \frac{1}{2}\log\det\left(S'(w)S(w)\right).$$

2.2 Inference

Set $\theta = (\sigma^2, w_1, \ldots, w_m, \alpha)$. We know that

$$\sqrt{n}\left(\hat{\theta} - \theta\right) \xrightarrow{d} N\left(0, IA\left(\theta\right)^{-1}\right),$$

where

$$IA\left(\theta\right) = \lim \frac{1}{n} I\left(\theta\right),$$

and $I\left(\theta\right)$ is the Fisher information matrix, given by

$$I\left(\theta\right) = -E_\theta\left(\frac{\partial^2 l\left(\theta\right)}{\partial\theta\partial\theta'}\right).$$

Now,

$$\frac{\partial l\left(\theta\right)}{\partial\theta_r} = \frac{1}{2} tr\left(H^{-1}\dot{H}_r\right) - \frac{1}{2}(y-\alpha 1)'\dot{H}_r(y-\alpha 1), r = 1, ..., m+1,$$

and

$$\frac{\partial^2 l\left(\theta\right)}{\partial\theta_r\partial\theta_s} = \frac{1}{2} tr\left(-H^{-1}\dot{H}_s H^{-1}\dot{H}_r + H^{-1}\ddot{H}_{rs}\right) - \frac{1}{2}(y-\alpha 1)'\ddot{H}_{rs}(y-\alpha 1), r, s = 1, ..., m+1.$$

Hence,

$$\begin{aligned} I_{r,s}\left(\theta\right) &= -\left[\frac{1}{2} tr\left(-H^{-1}\dot{H}_s H^{-1}\dot{H}_r + H^{-1}\ddot{H}_{rs}\right) - \frac{1}{2} tr\left(H^{-1}\ddot{H}_{rs}\right)\right] \\ &= \frac{1}{2} tr\left(H^{-1}\dot{H}_s H^{-1}\dot{H}_r\right), r, s = 1, ..., m+1. \end{aligned}$$

Also,

$$\frac{\partial l\left(\theta\right)}{\partial\alpha} = 1'H(y-\alpha 1)$$

$$\frac{\partial^2 l\left(\theta\right)}{\partial\alpha^2} = -1'H1 = -\frac{n}{\sigma^2}$$

and

$$\frac{\partial^2 l\left(\theta\right)}{\partial\alpha\partial\theta_r} = 1'\dot{H}_r(y-\alpha 1), r = 1, ..., m+1.$$

The asymptotic information matrix is seen to be

$$IA(\theta) = \begin{pmatrix} \left(\lim \frac{1}{2n} tr\left(H^{-1}\dot{H}_s H^{-1}\dot{H}_r\right)\right)_{1\leq r,s\leq m+1} & 0 \\ 0 & \frac{1}{\sigma^2} \end{pmatrix}.$$

A more explicit calculation of $IA(\theta)$ will be given in the next sub-section.

To conduct a hypothesis test of the form

$$H_0 : \theta_r = 0 \text{ vs. } H_1 : \theta_r > 0, r = 1, ..., m+1,$$

we need to use the statistic

$$\frac{\sqrt{n}\hat{\theta}_r}{\left(IA^{-1}\left(\hat{\theta}\right)\right)_{r,r}^{1/2}}.$$

Since the limit is generally unknown we can replace $IA\left(\hat{\theta}\right)$ by $I\left(\hat{\theta}\right)/n$ and use

$$t = \frac{\sqrt{n}\hat{\theta}_r}{\left(\left(I\left(\hat{\theta}\right)/n\right)^{-1}\right)_{r,r}^{1/2}} = \frac{\hat{\theta}_r}{\sqrt{\left(I^{-1}\left(\hat{\theta}\right)\right)_{r,r}}}. \tag{8}$$

We reject H_0 when t is large (e.g., when it exceeds 1.645, if a 5% significance level is desired).

Note that

$$\sqrt{n}\left(\hat{\alpha} - \alpha\right) \sim N\left(0, \sigma^2\right)$$

in finite samples. The variance σ^2 of $\sqrt{n}\left(\hat{\alpha} - \alpha\right)$ follows from the $(m+2, m+2)$-th element of the inverse of $IA(\theta)^{-1}$.

For multiple linear hypotheses of the form

$$H_0 : R\theta = r \text{ vs. } H_1 : R\theta \neq r,$$

where R is a $q \times (m+2)$ matrix consisting of $q < (m+2)$ independent linear hypotheses, we can use the Wald test, given by

$$W = \left(R\hat{\theta} - r\right)' \left[RI^{-1}\left(\hat{\theta}\right)R'\right]^{-1} \left(R\hat{\theta} - r\right).$$

The statistic is asymptotically distributed $\chi^2(q)$ under H_0. We reject H_0 when W is large.

2.3 Calculation of $IA(\theta)$

Some simplification of the calculation of $IA(\theta)$ results from the following.

$$\dot{H}_1 = -\frac{1}{\sigma^2}H$$

from which follows that

$$IA_{1,1} = \frac{1}{2\sigma^4}.$$

Also,

$$\dot{H}_r = \frac{\dot{S}_r'S + S'\dot{S}_r}{\sigma^2}, r = 2, ..., m+1$$

implying that

$$\begin{aligned}
\frac{1}{2n}tr\left(H^{-1}\dot{H}_rH^{-1}\dot{H}_1\right) &= -\frac{1}{2n\sigma^2}tr\left(H^{-1}\dot{H}_r\right) \\
&= -\frac{1}{2n\sigma^2}tr\left(\left(\frac{S'S}{\sigma^2}\right)^{-1}\frac{\dot{S}_r'S + S'\dot{S}_r}{\sigma^2}\right) \\
&= -\frac{1}{n\sigma^2}tr\left(S^{-1}\dot{S}_r\right).
\end{aligned}$$

3 Data and Results

3.1 Data

We obtained two databases of apartments, one consisting of apartments for sale, and one – for rent. Both databases are maintained by the Student Association of Tel-Aviv University.[3] Tel-Aviv University students have free access to the databases, whereas non-students can obtain it for a fee. Anyone may post an apartment in the appropriate database for a fee. Posting an apartment is done by filling out a questionnaire over the phone, where certain data are mandatory, and various verbal descriptions can be added as comments. Each posting is paid for two months, but it is updated every two weeks at most. At the end of a two-week cycle, the owner of the apartment is

[3] We thank the Student Association of Tel-Aviv University for the data.

called and asked whether she wishes to keep the posting, and if so, whether she would like to update the asking price. The database is therefore best conceptualized as atemporal: the asking price of an early posting may be updated in light of newer asking prices that were posted later on. This is reflected in the seemingly circular nature of the system (4).

The two databases were sampled at the same time, early August 2003. The rental database contained about 2000 entries, whereas the sales database – about 300. This size difference is typical because the students, who have free access to the databases, are more often interested in renting than in buying apartments.

All apartments were in the greater Tel-Aviv area. In more remote (and less expensive) suburbs there were mostly apartments for sale. To control for a possible effect of the suburb/township, we restricted attention to three municipalities, in all of which there were relatively large number of apartments in both databases: Tel-Aviv, Ramat-Gan, and Givataim. These municipalities are geographically contiguous.

Ideally, we would like to have the exact location of each apartment as part of the data. Unfortunately, the databases only contained street names, rather than exact addresses.[4] We therefore approximated the street address by the exact location of the midpoint of the street. We excluded from the data very long streets, for which such an approximation would not be very informative. We ended up with $n = 1240$ apartments for rent, and $n = 219$ apartments for sale.[5]

The complete list of variables for each database is given in Appendix A.

[4] This is typical of such databases. Because sellers normally do not grant real estate agents exclusivity rights, agents do not provide the exact address until they meet the buyer/renter and have them sign an exclusivity form. As a result, exact addresses almost never appear in public postings.

[5] We thank Professor Juval Portugali of Tel-Aviv University for access to a database that contained street lengths, as well as geographical coordinates of each street's midpoint.

3.2 Method

Each database was split two: a sample (learning database), consisting of 75% of the observations, and a prediction (test) database, consisting of the remaining 25%. The prediction database was selected as each fourth observation. Since the observations were ordered by the apartment size, the sample and prediction databases were slightly more representative of the entire database they were drawn from than a completely random selection would have been.

For the sales and the rental database we performed the following. (i) Regressing Y on $X^1, ..., X^m$ in the sample; (ii) finding the maximum likelihood similarity function for the system (4) in the sample; (iii) computing the maximum likelihood values for the two models (regression and similarity) on the sample; (iv) generating predictions for the prediction database using the two methods, and computing their SSPE (sum of squared prediction errors).

3.3 Results

Appendix B contains the estimated values of the relevant parameters and their standard deviations.

The main results are reported in Table 1.

Insert Table 1 about here

Table 1 reports the value of the log-likelihood function (LIKE) and the value of the sum of squared prediction errors (SSPE) for the two databases, for both the regression and the similarity models.

Table 1 shows that on the database of apartments for sale, the regression model performs better than does the similarity model: the likelihood function in the sample is higher for the regression, and the SSPE out-of-sample is

lower. This pattern is reversed in the database of apartments for rent: in this database, the similarity model achieves a higher value of the likelihood function, as well as lower value of the SSPE.

The results appear to support our hypothesis: in databases of apartments for sale, the rule-based (regression) model performs better than does the case-based (similarity) model, both in terms of maximizing the likelihood function in the sample and in terms of minimizing the sum of squared errors out-of-sample. This pattern is reversed in databases of apartments for rent.

3.4 Statistical Issues

Perusing Tables 1 and 2, one notices the difference in the sample size between the two databases considered. The number of apartments for sale, $n = 219$, is lower than the number of the apartments for rent, $n = 1240$, by a factor of 6 almost. This discrepancy raises the question, can the difference between the performance of the regression and the similarity models in the two database be simply due to the sizes? That is, is it possible that the effect we have found is solely a statistical artifact, and has nothing to do with the economic reasoning behind purchase and rental decisions?

This possibility might appear quite plausible. The regression model uses the data only for the estimation of the regression equation. If the data generating process (DGP) were indeed (1), and if we were to miraculously discover the actual parameters $\beta_0, \beta_1, ..., \beta_m, \sigma^2$, then we would need no further data in order to make the best predictions possible. By contrast, the similarity model (3) is inherently data-dependent. Datapoints are not only used to estimate the parameters $\alpha, w_1, ..., w_m, \sigma^2$: datapoints also enter the DGP of (4) itself. Hence, having a larger database will improve the predictions generated by the similarity model even if the true parameters were known to us. Conversely, more datapoints may improve the predictions of the similarity model even if the estimates of the parameters $w_1, ..., w_m$ are not accurate.

To see this point more clearly, assume that the actual DGP involves a

non-linear relationship between Y and $X^1, ..., X^m$. The regression model is restricted to linear relationships. By contrast, the similarity model generates predictions according to

$$\hat{Y}_{n+1} = \frac{\sum_{i \leq n} \hat{s}(X_i, X_{n+1}) Y_i}{\sum_{i \leq n} \hat{s}(X_i, X_{n+1})} \tag{9}$$

where $\hat{s}$ is the estimated similarity function. Thus, for every prediction $\hat{Y}_{n+1}$, the similarity model uses all datapoints, in a formula that may be viewed as local interpolation. This prediction is akin to the Nadaraya-Watson estimator for non-parametric regression, where the estimated $\hat{s} / \sum_{i \leq n} \hat{s}(X_i, X_{n+1})$ plays the role of the kernel function. Finding the appropriate kernel function is typically considered a theoretical problem. In our model, we turn it into an empirical problem.[6] But even a similarity function $\hat{s}$ that does not have the optimal weights $w_1, ..., w_m$ could serve as a kernel function, and may be expected to generate better predictions for Y than would linear regression, provided that n is large and that the similarity function is not too "flat".

To test the possibility that our results are solely an artifact of the sample size, we ran the two models on sub-samples of the rental database. The number of datapoints in the sample of the sales database was 164 (roughly 75% of $n = 219$). Hence we wished to test the models on a sub-sample of $n_k = 164$ datapoints from the rental database. Recall that the corresponding number in the entire rental database was 930 (75% of 1240). We also took a sample for an intermediate value of 620 (a half of 1240). For each sample size, $n_k = 164$, 620, and 930, we selected a sample of the apartments for rent, ran the two models, and compared them in terms of LIKE and SSPE. The SSPE was computed over the remaining database. Thus, for a sample of n_k datapoints we had a prediction database containing $(1240 - n_k)$ observations. The results are reported in Table 2.

[6] Finding an optimal bandwidth for the kernel function is often done empirically. In our model, all m parameters of the kernel functions are estimated from the data, allowing us to empirically determine their relative importance.

Insert Table 2 about here

Table 2 indicates that the sample size does indeed have an effect on the relative performance of the two methods. Considering the LIKE criterion first, the similarity model does not perform as well as the regression model for a small sample ($n_k = 164$). The two models have very similar likelihood values for a mid-size sample ($n_k = 620$), and it is only for a large sample ($n_k = 930$) that the similarity model performs better than does the regression model.

Turning to the SSPE criterion, it turns out that the similarity model performs better than does the regression model on all three databases. Yet, when we compare the SSPE's generated by the two models, we find that for a larger sample the advantage of the similarity model increases. To see this, we computed the ratio of the SSPE of the regression to the SSPE of the similarity model (in the last column of the table). As can be seen, this ratio grows with n_k: whereas the regression model's prediction is worse than that of the similarity model only by 11% for a small sample, this factor grows to 28% for a large database.

Thus, our data indicate that the statistical effect we suspected does indeed exist. Yet, it is important to note that this statistical effect does not explain the entire pattern of results obtained. Even for a small database, the SSPE of the similarity model was lower than that of the regression model, while this pattern was reversed on an equally-sized database of apartments for sale. Hence, the statistical effect cannot be solely responsible for the results reported in Tables 1 and 2, and the economic effect we hypothesized probably plays a role as well.

Table 2 also suggests that if we had a larger database of apartments for sale, it is quite possible that the similarity model would have obtained better results than would the regression model. Generally speaking, one

should expect the similarity model to perform better for larger databases. We conjecture that this statistical effect would be independent of the type of the data analyzed. The economic effect, however, implies that for rental data the similarity model would be better than the regression model already for smaller databases than for sales data.

The statistical effect we conjecture might also be reflected in human reasoning. Specifically, it is possible that people use rule-based reasoning when they have a database that is not too large, but that they switch to case-based reasoning when the database is very large. This might be optimal because, when the database is large enough, there is no need to develop theories (or rules): every possible instance, that is, every relevant combination of values of $X^1, ..., X^m$, has enough cases in memory that are similar to it, for the person to be able to come up with a good assessment of the value of Y based on these similar cases.[7]

Observe that the similarity model performs better than the regression model in terms of a low SSPE already for small sample sizes (low values of n_k above), whereas a better performance in terms of a higher LIKE is obtained only for larger samples. We speculate that this pattern is not coincidental. The reason might be the following. The LIKE criterion is the criterion by which we choose the parameters of both models. It should therefore be expected that the parameters chosen for a particular sample will not perform as well on the prediction database (out-of-sample).[8] This bias exists to the

[7] When the database is very small, it may not contain enough datapoints to support any theory. Thus, case-based reasoning may be more prevalent than rule-based reasoning for small and for large databases, whereas rule-based reasoning may be more prevalent for medium-sized databases, that contain enough observations to generate theories, but not enough observations to do without theories.

[8] This might be viewed as a type of "regression to the mean" phenomenon: the particular values of the parameters that we choose are those that happen to perform well in the sample. Part of the success of these parameters might be due to random factors, and these need not be equally auspicious outside the sample. It follows that one should not expect the chosen parameters to perform on a new database as well as they did on the sample.

same degree for the regression and for the similarity model. However, the similarity model has a self-correction mechanism: because it uses the entire database for each prediction it generates, it may perform well out-of-sample even if the similarity function, which was chosen based on in-sample performance, is not necessarily the best one. By contrast, the regression model does not have any similar self-correction mechanism: the regression coefficients that were chosen based on their in-sample performance are used for out-of-sample prediction with no further aid from the data.

4 Concluding Remarks

It stands to reason that certain combinations of the regression and the similarity models may perform better than both in terms of providing the best fit. For instance, one may use our similarity-weighted average and plug it into the regression model as another explanatory variable. This would resemble a hedonic spatial regression, in which one attempts to estimate the weight matrix (along the lines suggested in this paper). However, such a hybrid model will not be able to compare the two modes of reasoning in their pure form.

We do not expect to obtain a qualitatively clear result, saying that people think in terms of cases or in terms of rules. We believe that both modes are involved in almost any reasoning, and that a variety of factors may affect their relative importance. Our focus in this paper is on a particular economic factor, namely, the nature of the market under discussion. We conjecture that in general, in comparison to rule-based reasoning, case-based reasoning will be more prevalent in non-speculative markets than in speculative ones.

5 Appendix A: The Variables

Insert Table 3a about here

Insert Table 3b about here

6 Appendix B: Estimates of Parameters

Insert Table 4a about here

Insert Table 4b about here

References

[1] Anselin, L. (1988), Spatial Econometrics: Methods and Models, Dordrecht: Kluwer.

[2] Brasington and Hite (2004), *Regional Science and Urban Economics*, .

[3] Dubin, R. (1988), "Spatial Autocorrelation", Review of Economics and Statistics, 70, 466-474.

[4] Gao, X., Y., Asami, and C. J. Chung (2002), "An Empirical Evaluation of Hedonic Regression Models", Symposium on Geospatial Theory, Processing, and Applications.

[5] Kim, C. W., T. T. Phipps, and L. Anselin (2003), "Measuring the Benefits of Air Quality Improvements: A Spatial Hedonic Approach", *Journal of Environmental Economics and Management*, **45**: 24-39.

[6] Ord, K. (1975), "Estimation Methods for Models of Spatial Interaction", *Journal of the American Statistical Association*, **70**: 120-126.

[7] Ripley, B. (1981), Spatial Statistics, New York: John Wiley.

[8] Ripley, B. (1988), Statistical Inference for Spatial Processes, Cambridge: Cambridge University Press.

[9] Rosen, S. (1974), "Hedonic Prices and Implicit Markets: Product Differentiation in Pure Competition", *Journal of Political Economy*, **82**: 34-55.

Table 1: LIKE and SSPE, for regression and similarity, and for the two databases.

	Sales ($n = 219$)		Rent ($n = 1240$)	
	Regression	Similarity	Regression	Similarity
LIKE	-876	-902	$-5,420$	$-5,380$
SSPE	$146,759$	$185,160$	$2,550,834$	$1,985,600$

LIKE – Value of the log-likelihood function (in-sample, 75% of the data points)

SSPE – Sum of Squared Prediction Errors (out of sample, remaining 25% of the data points)

Table 2: LIKE and SSPE for the two models, for various samples of the rental database.

k		Regression	Similarity	SSE ratio
164	MLE	-951	-974	
	SSE	8,690,667	7,853,000	1.11
620	MLE	-3630	-3,605	
	SSE	4,543,413	3,783,500	1.20
930	MLE	-5,420	-5,380	
	SSE	2,550,834	1,985,600	1.28

LIKE – Value of the log-likelihood function (in-sample, 75% of the data points)

SSPE – Sum of Squared Prediction Errors (out of sample, remaining 25% of the data points)

Table 3a: Variables Names and Descriptions – Sales database

Variable	Description
Rooms	
Size	in m^2
Floor	
Elevator	indicator
Parking	indicator
Air-conditioning	indicator
Renovated	verbal
Quiet	verbal
Balcony	verbal
x coordinate	
y coordinate	
No sections in the street	indicates length of street
View	verbal
Roof	verbal
Direction of ventilation	no. of directions of apt's windows
Face front	indicator
Face rear	indicator
Face both	indicator
Other	

Comments (for Tables 4a and 4b): "Indicator" variables are mandatory. "Verbal" variables are also indicator variables that were picked from the verbal description. The variables "x coordinate", "y coordinate", and "No of sections in the street" were obtained from the geographical database using the street name. The rest of the variables originate from the posting.

Table 3b: Variables names and descriptions – Rentals database

Variable	Description
Rooms	
Big	verbal
Floor	
Elevator	indicator
Parking	indicator
Air-Conditioned	indicator
Renovated	verbal
Quiet	verbal
Balcony	verbal
x coordinate	
y coordinate	
No sections in the street	indicates length of street
Furnished	verbal
Garden	verbal
Duplex	verbal
Gallery	indicates a sleeping gallery (loft style)
Studio	verbal
Washer	verbal
Boiler	verbal
Villa	verbal
Roof	verbal
Bars	indicates if windows have bars

(See Comments following Table 3b.)

Table 4a: Variables and estimated coefficients (and standard deviations) for Regression and Similarity – Sales database

Variable	Regression $\hat{\beta}_j$	Similarity $\hat{w}_j$
Rooms	14.163	453.650*
	(8.420)	(21.334)
Size	1.625*	0.002*
	(0.293)	(0.000)
Floor	−2.285	0.000
	(3.018)	(0.012)
Elevator	24.234*	0.000
	(11.800)	(0.201)
Parking	5.898	0.638*
	(10.426)	(0.244)
Air-conditioning	8.869	0.000
	(10.785)	(0.212)
Renovated	4.093	0.000
	(8.771)	(0.200)
Quiet	19.210	0.000
	(10.505)	(0.127)
Balcony	0.662	0.000
	(10.944)	(0.189)
x coordinate	−0.008*	0.000*
	(0.003)	(0.000)
y coordinate	0.000	0.000
	(0.000)	(0.000)

Variable	Regression $\hat{\beta}_j$	Similarity $\hat{w}_j$
No sections in the street	-1.060	0.342^*
	(0.915)	(0.014)
View	21.184	86.016^*
	(15.434)	(10.668)
Roof	18.376	0.000
	(22.229)	(0.356)
Direction of ventilation	6.322	0.000
	(15.143)	(0.204)
Face front	10.988	0.000
	(12.108)	(0.213)
Face rear		0.260
		(0.185)
Face both	18.587	0.000
	(13.030)	(0.173)
Other	3.106	0.094
	(12.595)	(0.196)
C	$1,388.505^*$	192.415^*
	(474.029)	(4.286)

$*$ – Significant at the 5% level.

Standard deviation of 0.000 indicates a positive number smaller than 0.0005.

Table 4b: Variables and estimated coefficients (and standard deviations) for Regression and Similarity – Rentals database

Variable	Regression $\hat{\beta}_j$	Similarity $\hat{w}_j$
Rooms	127.033*	3,496.100*
	(4.247)	(360.011)
Big	14.821*	0.806*
	(6.148)	(0.076)
Floor	12.836*	0.226*
	(2.225)	(0.016)
Elevator	36.733*	0.000
	(10.644)	(0.087)
Parking	16.513	0.000
	(8.966)	(0.099)
Air-Conditioned	25.677*	25.460*
	(6.604)	(0.934)
Renovated	7.519	0.000
	(5.734)	(0.064)
Quiet	1.940	0.000
	(5.776)	(0.067)
Balcony	19.356*	3.636*
	(6.377)	(0.157)
x coordinate	−0.017*	0.000*
	(0.002)	(0.000)
y coordinate	0.000*	0.000
	(0.000)	(0.000)

Variable	Regression $\hat{\beta}_j$	Similarity $\hat{w}_j$
No sections in the street	-1.533^*	0.009^*
	(0.530)	(0.001)
Furnished	16.365^*	0.000
	(5.984)	(0.074)
Garden	20.339	6.385^*
	(13.653)	(0.520)
Duplex	40.773	82.656^*
	(48.352)	(20.846)
Gallery	-19.548	0.000
	(24.427)	(0.063)
Studio	37.977^*	11.847^*
	(17.943)	(0.374)
Washer	-20.661	0.052
	(22.659)	(0.292)
Boiler	3.610	0.192^*
	(7.267)	(0.091)
Villa	11.619	0.000
	(32.616)	(0.796)
Roof	-10.825	0.932^*
	(20.419)	(0.133)
Bars	24.676^*	0.000
	(7.176)	(0.059)
C	$3,298.435^*$	593.090^*
	(297.705)	(2.494)

$*$ – Significant at the 5% level.

Standard deviation of 0.000 indicates a positive number smaller than 0.0005.