



*The World's Largest Open Access Agricultural & Applied Economics Digital Library*

**This document is discoverable and free to researchers across the globe due to the work of AgEcon Search.**

**Help ensure our sustainability.**

Give to AgEcon Search

AgEcon Search

<http://ageconsearch.umn.edu>

[aesearch@umn.edu](mailto:aesearch@umn.edu)

*Papers downloaded from **AgEcon Search** may be used for non-commercial purposes and personal study only. No other use, including posting to another Internet site, is permitted without permission from the copyright owner (not AgEcon Search), or as allowed under the provisions of Fair Use, U.S. Copyright Act, Title 17 U.S.C.*

*No endorsement of AgEcon Search or its fundraising activities by the author(s) of the following work or their employer(s) is intended or implied.*

# **Identifying Submarkets in the Wine Industry: a Multivariate Approach to Hedonic Regression**

Authors:

**Marco Costanigro**

School of Economic Sciences  
Washington State University  
PO Box 646210, Pullman, WA, 99164  
Phone: (509) 335-8600  
Email: costanigro@yahoo.com

**Jill McCluskey**

School of Economic Sciences  
Washington State University

**Ron Mittelhammer**

School of Economic Sciences  
Washington State University

## **Working Paper**

*Selected Paper prepared for presentation at the American Agricultural Economics Association Annual Meeting, Long Beach, California, July 23-26, 2006*

*Copyright 2006 by Marco Costanigro Jill McCluskey and Ron Mittelhammer. All rights reserved . Readers may make verbatim copies of this document for non-commercial purposes by any means, provided that this copyright notice appears on all such copies.*

## **Identifying Submarkets in the Wine Industry: a Multivariate Approach to Hedonic Regression**

Marco Costanigro<sup>1</sup>, Jill McCluskey<sup>2</sup> and Ron Mittelhammer<sup>3</sup>

Several authors have utilized the hedonic approach (Rosen, 1974) to investigate the determinants of wine prices. Most of the research effort has been directed so far to determine which attributes are good candidates as explanatory variables in the hedonic function. Even though results in this area have been constrained by the nature of the available data, there is substantial agreement about what influences wine prices.

Combris et al. (1997, 2000) showed that when regressing objective characteristics and sensory characteristics on wine price, the objective cues (such as expert score and vintage) are significant, while sensory variables such as tannins content and other measurable chemicals are not. Nevertheless, substantial evidence (Oczkowski 1994; Landon and Smith 1997; Shamel *et al.* 2003, Angulo *et al.* 2000) indicates that ratings by specialized magazines are significant and should be included in the hedonic function when modeling wine prices. Possible explanations for the insignificance of sensory cues are the difficulty of isolating the effect of each chemical on the final flavor and smell and that only a small percentage of wine purchasers are connoisseurs. Therefore, expert ratings act as a signal to the consumer. It is uncertain whether expert ratings influence prices because they are good proxies for quality of the wine or because of their marketing effect. In addition to expert ratings, the region of production, capturing the effects of the collective reputation of the district, and the vintage are often reported as significant variables (Angulo *et al.*, 2000; Schamel and Anderson, 2003).

---

<sup>1</sup>Ph.D. candidate, School of Economic Sciences, Washington State University.

<sup>2</sup>Associate Professor, School of Economic Sciences, Washington State University.

<sup>3</sup>Regent Professor and Chair, School of Economic Sciences, Washington State University

Marketing wine research focuses more on the behavioral aspect of wine purchasing. Spawton (1991) identifies four different categories of wine consumers: connoisseurs, aspirational drinkers, beverage wine consumers and new wine drinkers. Each buyer type has different attitudes and preferences relating to wine. For all types, the main factors influencing the purchasing decision are previous experience and knowledge of the product, objective cues such as production region, brand, and label, the occasion in which the wine will be consumed, and the price itself.

Hall *et al.* (2001) highlighted the relationship between product choice and occasion of consumption. Their findings suggest that consumers look for different attributes, or value the same attributes differently, depending on the occasion in which the wine is meant to be consumed. Also, they show that price is often considered a quality cue, helping consumers to associate wine and occasion of consumption.

In his seminal work, Rosen (1974) developed the hedonic framework in the context of a given product class, where a product class is a set of goods that are somewhat differentiated, but are so similar that consumers consider them as variations of the same product. Recent work (Costanigro *et al.*, under review) provided empirical evidence that implicit prices of several wine attributes significantly vary across price ranges, implying that the traditional approach of hedonic regression over the pooled price range produces biased estimates. This suggests that wine is a composite product class, and that smaller subclasses exist within it. Differences in implicit prices involved most of the attributes. In particular, certain regions of production were shown to award price premia in the cheaper wine classes, while they have no effect for the more expensive

ones. Furthermore, the effect of cellaring time was found to be very different across price segments, as one would intuitively expect.

In their work, the authors identified four market segments (cheap or commercial wines 1\$-13\$, semi-premium: 13\$-21\$, premium: 21\$-40\$ and ultra-premium: 40\$ and over) by selecting the three price-breakpoints that minimize the sum of squares errors of the overall model (SSE min or price range criterion in the remaining of this paper). This wine classification was found to fit surprisingly well with categorization used in the wine industry and based on the common knowledge of wine producers, as Ernst and Young Consulting (1999) explain in a report on the state of the Australian wine market.

Clearly, the identification of the wine subclasses is a very fundamental step that strongly influences results and the limitation of the price range approach is that it assumes that price contains all the information necessary to separate wine classes. Ideally, wines residing in the same product class should possess more homogeneous characteristics in the whole vector of wine attributes, and not exclusively price. This suggests a multivariate approach to the problem of determining wine classes, which would allow for a more holistic determination of product classes.

An interesting parallel with the problem at hand can be drawn from the issue of market segmentation in the housing hedonic literature. Straszheim (1974) first argued that it is appropriate to segment markets when analyzing property values. He showed that by estimating separate hedonic price functions for different geographic areas of the San Francisco Bay area, the sum of squared errors in predicting prices across the entire sample was significantly reduced. Market segments and product classes are not necessarily synonyms, but they share the same idea that the more two products are

differentiated (we could think of location as an attribute), the less they are fungible. So as the vector of attributes of two products diverges, the use that consumers make of them diverges too. This transition can be gradual, perhaps generating hybrid products that can be employed for multiple purposes, or the presence/absence of an attribute might unambiguously determine the membership/non membership of a good to a given product class. Also, the cost of assembling a given bundle of attributes in the same product will change as we change the vector of attributes. This has straightforward implications for the implicit prices: as use and production processes of two similar goods diverge, the market valuation of the attributes will diverge too, even for the attributes that are common to both products. The cost of ignoring this process is biased estimated coefficients of the hedonic price function.

The problem of multiple product classes can be solved with two alternative (but dual) approaches: the first is the one suggested by Straszheim (1974), which consists of trying to find a criterion to partition the data and estimate simpler hedonic functions specific to a sub-sample. The second approach tries to formally model the existence of product classes on the right-hand side, specifying hedonic equations with enough flexibility to allow for different parameters for products in different subclasses. Examples of this methodology are random parameter models (see Allenby and Rossi, 1999 for an example) and latent class models (see Greene 2001 for a survey of the existing literature). The drawback of this approach is that large, high quality datasets are required and the resulting models are complex and hard to interpret.

Most researchers in the real estate hedonic literature chose to face the problem of market segmentation with the data partition approach and use multivariate analysis to

partition the data as an alternative to using administrative boundaries or zip code areas. The general idea is to produce clusters having characteristics that are homogeneous under a given criterion. Dale-Johnson (1982) uses Q-factor analysis to assign data points to a set of clusters characterized by a number of representative transactions. He then estimate hedonic models specific to the identified data clusters. Bourassa *et al.* (1999) uses principal component analysis to extract factor scores of each observation. Then, he applies cluster analysis (K-means and Ward clustering) on the factor scores to determine the composition of the submarkets. The final step is, again, estimation of hedonic functions specific to the market segment. In a later paper, Bourassa *et al.* (2003) test their technique in out-of-sample prediction performance, and conclude that market segmentation based on administrative boundaries used by appraisers is more effective than data driven clustering of the data. Similarly, Watkins (1999) uses a two stage approach in which he first identifies homogeneous submarkets by principal component factor analysis, and then estimates the submarket hedonic functions. A simpler approach is proposed by Wilhelmsson (2004), who identifies submarkets using Ward clustering of the OLS residuals from a regression performed using the whole dataset, therefore ignoring submarkets. He shows that the within sample prediction ability of his model is superior to a segmentation based on administrative areas. Also, he shows that spatial dependency of the error term is reduced by the method.

Goodman and Thibodeau (1998) model market segmentation on the right-hand side. They do this using a two stage hierarchical model and defining submarkets as the areas in which the per-unit price of housing is homogeneous. In their paper market segmentation is assumed to be driven by the quality of public schools in the

neighborhood so that the data is supplemented by elementary school student performance. The resulting model is complex and involves two steps: expectation-maximum likelihood estimation and a maximum likelihood step.

In this paper we avoid a formal modeling of the existence of product class specific estimated coefficients, and investigate the effectiveness of the data partitioning approach at identifying classes. The objective of which is to produce wine class specific estimates of the implicit prices of wine attributes. The methodology adopted is therefore one of product class specific hedonic functions: drawing from the existing literature we use several multivariate analysis techniques to produce alternative clustering of the data. The clustering techniques are evaluated on the basis of their ability to produce identifiable clusters and out-of-sample prediction performance of the resulting models. Section 2 presents in detail the data, methods and the empirical model being used. In section 3 we present and discuss the results, highlighting the shortcomings of multivariate cluster analysis in the context of the paper. At the end of the paper we propose a promising alternative method involving local regression and cluster analysis that takes a step towards a formal modeling of the existence of wine classes, but retains the simple framework of the multivariate approach.

## 2. Data, Methods and Empirical Model

### Data

The data set is composed of 13,157 observations derived from ten years (1991-2000) of tasting ratings reported in the *Wine Spectator* Magazine (online version) for California (11,869 observations) and Washington (1,288 observations) red wines. For the purpose of out-of-sample testing the dataset was randomly divided in two: a working sample containing about 75% of the data in which models are estimated and a testing sample containing the remaining observations, which is used for prediction.

Four of the variables are non-binary: price of the wine adjusted to 2000 values by a consumer price index (CPI) for alcohol, score obtained in the expert sensory evaluation provided by the Wine Spectator, the number of cases produced, and the years of aging before commercialization. Descriptive statistics for these variables are reported in table 1. Note that wine prices have a skewed distribution, but the majority of the observations fall in the \$10 to \$50 range. Indicator variables were used to denote regions of production, wine varieties, and the presence of label information. The regions of production for California wines include Napa Valley, Bay Area, Sonoma, South Coast, Carneros, Sierra-Foothills and Mendocino, while Washington wines were not separated by regions. These geographical partitions are the ones adopted by the Wine Spectator to categorize the wines, often pooling several American Viticultural Areas (AVAs) in the same region. Varieties include Zinfandel, Pinot, Cabernet, Merlot and Syrah grapes, as well as wines made from blending of different varieties (non-varietal). The vintage year is available for each wine along with other label information such as “reserve” and

“estate produced.” Table 2 reports all variables and abbreviations used throughout the paper with a short description.

**Table 1:** descriptive statistics of quantitative explanatory variables

	Variable							
	California				Washington			
	Price*	Cases	Score	Age	Price*	Cases	Score	Age
N	11869	11869	11869	11869	1288	1288	1288	1288
Mean	31.06	6719	86.115	2.7646	23.262	6720	86.815	2.8346
St. Dev	51.44	26201	3.955	0.7429	12.523	30764	3.38	0.7714
Median	22	1467	87	3	20	1000	87	3
First Quartile	15	500	84	2	5	377	85	2
Third Quartile	35	6000	88	3	144	2638	89	3
Minimum	3	16	60	1	5	45	67	1
Maximum	2000	950000	99	9	144	550000	96	7

\*Adjusted by a CPI index for alcohol

**Table 2:** short descriptions of the abbreviation used for the explanatory variables

Predictor	Short Description
Score	Rating Score from the Wine Spectator
Score^2	Score Squared
Age	Years of Aging Before Commercialization
Age^2	Age Squared
Cases	Number of Cases Produced
Ln(cases)	Natural Log of Hundreds of Cases Produced
Napa	Region of Production
Bay area	
Sonoma	
South coast	
Carneros	
Sierra foothills	
Mendocino	
Washington	
Nonvarietal	
Pinot noir	
Cabernet	
Merlot	
Syrah	
Reserve	"Reserve" was Reported on the Label
Vineyard	Specific Name of the Vineyard on the Label
Estate	"Estate" Produced Wine
91, ..., 99	Vintage
Wa	Washington State wines

## Methods

### Ward and K-means clustering algorithms

To identify homogeneous wine classes, two data clustering techniques were used: Ward and K-means clustering. Ward is a hierarchical clustering algorithm that minimizes the sum of squared deviations ESS from the cluster centroid, given a desired final number of clusters. For each cluster:

$$2.1. \quad ESS_i = \sum_{j=1}^N (x_j - \bar{x})'(x_j - \bar{x})$$

where  $i$  indexes clusters ( $i=1, \dots, k$ ),  $x_j$  is the multivariate measurement associated with the  $j_{th}$  observation and  $\bar{x}$  is the mean of all items in the cluster. In the first step each observation is considered as a cluster. At each iteration all the possible unions of the existing clusters are considered, and then the union that minimizes the sum of ESS over all clusters is implemented. The algorithm continues aggregating until a desired number of clusters, say  $k$ , are obtained.

K-means is a non-hierarchical technique that starts from *a-priori* assignment of the items to a predefined number of clusters, and then iteratively reassigns each item to the existing cluster whose centroid is nearest. It is known that the beginning *a-priori* partition for the first iteration can highly influence the final clustering when using K-means, so good starting partitions are advised.

For both algorithms the final number of cluster was set to  $k=4$ , mimicking the number of wine classes suggested by the Ernst and Young report (1999). For the K-means approach the partition of the data based on the price range approach of Costanigro *et al.* (under review) was used as initial clustering for the algorithm.

## **Alternative Approaches**

Ward and K-means clustering was performed on several variables, obtaining alternative partitions in four groups of the dataset. A first approach included all the information included in the dataset, therefore clustering on price, tasting score, cellaring time, number of cases produced, grapes variety, macro-region of production, vintage and other information reported in the label. Obviously this approach includes both binary and non-binary variables. It should be noticed that the presence of several binary variables, a feature common to most hedonic models, complicates the process of calculating distances across observations for clustering purposes. While standardizing easily solves the problem of having variables measured in different units, it does not change the fact that the distance between an item possessing a given qualitative attribute and an item that does not possess it is forced to be equal to one. Using the standardization only scales the distance to another scalar.

A second method involved using only the non-binary variables, namely price of the wine, aging, tasting score and cases produced. While this approach excluded a considerable amount of the information contained in the dataset, it seems reasonable that the mentioned variables carry enough information to identify wine classes. Furthermore, a very crucial aspect of data clustering is that all the variables introduced in the algorithm are equally weighted, and they contribute evenly to the calculated distance of two observations. This calls the researchers to carefully select the variables, as an improper choice will yield chaotic clusters.

The third method is a variation of Wilhelmsson (2004) approach, and consists of clustering on the fitted values and residuals obtained from a hedonic model estimated for

the whole dataset, ignoring product subclasses. Lastly, we extract a number of principal components from the non-binary variables (correlation matrix) and construct a set of orthogonal factor scores variables explaining at least 80% of the variation of the non-binary variables as in Bourassa *et al.* (2003) and Watkins(1999).

A schematization of the steps common to all approaches is as follows:

- 1) **Working sample:** obtain a partition of the data in four clusters using one of the introduced methods, then estimate hedonic model specific to each data cluster via OLS
- 2) **Testing sample:** obtain four clusters of data using the same method used in 1, and then match each cluster in the testing sample with the equivalent cluster of the working sample. Once cluster are properly labeled, use the model estimated in the working sample to predict wine prices in the testing sample. The prediction performance was then evaluated using a Median Percent Error Rate (MPER), calculated as:

$$\text{median}[abs(y - \hat{y}) / y]$$

### **Evaluating alternative approaches**

A crucial requirement for all the proposed techniques is the ability to produce clusters that are consistent across working and testing samples and also identifiable as a specific wine class. The need for robustness of results across working and testing samples is dictated by the necessity to use the correct model to predict out-of-sample. The second requirement follows from the fact that the main objective of this paper is to estimate wine-class specific implicit prices of the attributes that can provide information useful to consumers and producers. This point is not considered as important in the housing hedonic literature, where the primary concern is price prediction for value appraisal

purposes. In this work a lack of cluster identification would make the analysis a mere econometric exercise.

An easy way to label clusters is to examine the within cluster price distribution. Intuitively, we expect the interquartile range of the price distribution for a commercial wine to span price values that are lower than an ultra premium wine. Therefore, we use the within cluster median price to label the data partitions resulting from the different approaches as wine classes.

### Empirical Model

Hedonic theory does not provide a particular specification of the functional form for regressing price on the attributes. Models are usually empirically designed, and therefore flexible functional forms are desirable. Specification tests and the conformity with assumption of OLS regression further drive the work of the researchers. The functional form used in this work was derived in an earlier paper (Costanigro *et al.*, under review), using the whole dataset, and is the following:

$$\begin{aligned}
 2.2. \quad Price^{-0.5} = & \left( \beta_1 + \beta_1^w WA \right) (Score) + \left( \beta_2 + \beta_2^w WA \right) (Score)^2 + \left( \beta_3 + \beta_3^w WA \right) (Age) + \left( \beta_4 + \beta_4^w WA \right) (Age)^2 \\
 & + \left( \beta_5 + \beta_5^w WA \right) LN(Cases) + \sum_{i=1}^5 \left( \beta_{5+i} + \beta_{5+i}^w WA \right) (Variety_i) + \sum_{i=1}^9 \left( \beta_{10+i} + \beta_{10+i}^w WA \right) (Vintage_i) \\
 & + \sum_{i=1}^3 \left( \beta_{19+i} + \beta_{19+i}^w WA \right) (Label_i) + \sum_{i=1}^7 \beta_{22+i} (Region_i) + \varepsilon_i
 \end{aligned}$$

We maintain the assumption that the adopted specification is valid for all the different partitions of the dataset. While this might seem quite a strong assumption, the transformation of the dependent variable was consistently effective in producing normally distributed residuals and the polynomial specification with intercept and slope shifters impose few constraints on the estimated hedonic price function.

### **3. Results and Discussion**

The clustering approach involving all the variables (including binary and non-binary) yielded data partitions that were not clearly identifiable as wine classes using the median price criterion. Also, clusters were not robust from working sample to testing sample. This makes prediction impossible (and useless) and thus the approach was abandoned.

Table 3 reports the number of observations in each cluster identified by the Ward and K-means algorithm, along with the median within-cluster price for each of the remaining approaches. Ideally for the same approach and the same clustering algorithm the median price of each partition should be clearly differentiated from the adjacent cluster (horizontally in the table), and as close as possible in the working and testing sample (vertically in the table). Obviously the SSE min approach, based solely on price ranges, yielded the most consistent data partitioning across working and testing sample and most clearly differentiated clusters within each dataset.

Overall, it was possible to label all the identified clusters as a wine class. The main concern is a lack of robustness in median price from working sample to testing sample, especially in the ultra premium class, where observations are sparser. In particular the K-means algorithm yielded the least robust results, and also showed a tendency to isolate outliers in clusters of very small size.

The performance of the different approaches in out-of sample prediction is reported in table 4. For each approach, we present the within sample  $R^2$ , a MPER calculated pooling the prediction for all cluster, and then a MPER calculated within each cluster. It is evident that the SSE min approach has the best performance, with a 10.9% overall MPER. The worst performing model is the pooled approach, which ignores

product subclasses and estimates a single model for all wines. The fact that the SSE min approach outperforms considerably the pooled approach in out-of-sample prediction, and not only in  $R^2$ , confirms that ignoring wine classes biases the parameters estimated in the pooled approach. Therefore estimating a price-range specific hedonic function is not a mere data overfitting.

In general the clustering approaches do slightly better than the model estimated for the pooled dataset (with the exception of Ward clustering on principal component factor scores), but are not comparable to the SSE minimization criterion. Ward clustering on the fitted and predicted values from the pooled model yielded the best results across all the clustering approaches, with an overall median percentage error of 17.7%. The performance of the approaches involving clustering is even more disappointing if we consider that these models estimates four times as many parameters as the pooled approach. Conversely, the  $R^2$  performance of these methods is quite reasonable, suggesting that the lack of out-of-sample predictive ability might be due to poor robustness of the clustering algorithms in the transition from working to testing sample.

**Table 3:** number of observation and median within-cluster price, for all the clustering approaches

			No classification		Commercial		Semi-Premium		Premium		Ultra Premium	
			N	Median Price	N	Median Price	N	Median Price	N	Median Price	N	Median Price
sample												
<b>Pooled</b>			9,890	\$23.00								
<b>SSE Min</b>		working			1,242	\$10.26	3,103	\$17.10	3,663	\$28.00	1,882	\$51.50
		testing			402	\$10.17	1,046	\$17.07	1,199	\$28.00	618	\$50.50
<b>Original Data</b>	k-means	working			56	\$9.27	3,673	\$16.95	6,142	\$28.25	19	\$631.00
		testing			7	\$7.98	1,221	\$16.35	1,859	\$27.25	178	\$85.25
	Ward	working			949	\$15.96	3,066	\$16.95	4,189	\$26.00	1,686	\$40.54
		testing			337	\$13.52	832	\$16.08	1,248	\$25.00	848	\$42.00
<b>Fitted/Residuals</b>	k-means	working			8,372	\$20.60	1,478	\$50.00	26	\$100.00	14	\$884.00
		testing			2,164	\$18.08	949	\$38.00	147	\$65.00	4	\$535.00
	Ward	working			4,921	\$17.00	3,080	\$25.25	1,346	\$42.42	543	\$75.92
		testing			1,739	\$17.00	871	\$26.16	247	\$49.00	408	\$50.50
<b>Principal Component</b>	k-means	working			56	\$9.27	3,542	\$19.62	4,676	\$22.00	1,616	\$50.00
		testing			8	\$8.49	1,203	\$18.72	1,447	\$21.00	607	\$48.48
	Ward	working			3,630	\$19.62	3,523	\$20.60	1,282	\$29.12	1,455	\$45.50
		testing			1,257	\$18.54	547	\$16.48	409	\$30.00	1,052	\$32.00

**Table 4:** out-of-sample prediction performance.

		Overall		Commercial		Semi-premium		Premium		Ultra premium	
		R <sup>2</sup>	MPER*	R <sup>2</sup>	MPER*	R <sup>2</sup>	MPER*	R <sup>2</sup>	MPER*	R <sup>2</sup>	MPER*
<b>Pooled</b>		69.40%	<b>18.90%</b>								
<b>SSE Min</b>		91.00%	<b>10.90%</b>	28.40%	<b>10.20%</b>	22.40%	<b>8.60%</b>	19.30%	<b>11.40%</b>	19.30%	<b>19.00%</b>
<b>Original Data</b>	k-means		<b>18.50%</b>	**	**	61.00%	<b>18.80%</b>	59.60%	<b>18.14%</b>	30.80%	<b>83.00%</b>
	Ward		<b>18.07%</b>	66.70%	<b>18.30%</b>	58.80%	<b>17.69%</b>	53.00%	<b>16.31%</b>	54.00%	<b>21.30%</b>
<b>Fitted/Residuals</b>	k-means		<b>18.80%</b>	61.30%	<b>17.30%</b>	26.50%	<b>23.60%</b>	**	**	**	**
	Ward		<b>17.70%</b>	62.00%	<b>15.90%</b>	56.70%	<b>16.10%</b>	52.80%	<b>16.10%</b>	61.20%	<b>39.40%</b>
<b>Principal Component</b>	k-means		<b>18.30%</b>	65.70%	<b>13.20%</b>	69.50%	<b>17.20%</b>	55.30%	<b>17.43%</b>	38.10%	<b>23.10%</b>
	Ward		22.01%	70.50%	<b>17.38%</b>	55.90%	<b>17.03%</b>	60.80%	<b>22.50%</b>	47.80%	<b>37.82%</b>

\* Median % error rate in out-of-sample prediction

\*\* Cluster is too small to estimate model

The poor performance of the clustering approaches can be explained considering several factors. We already mentioned the issues involved with mixed binary and non-binary data and how clustering algorithms consider every variable equally important in determining distances. Thus it is necessary that researchers have some *a priori* knowledge about which variables are the most likely to be linked with the characteristics that drive the change in product classes and implicit prices of the attributes. If this expertise is lacking, the resulting clusters will be likely to be chaotic and unidentifiable. Furthermore, there will be no real theoretical basis to expect that the clustering algorithms will result in data partitions in which the coefficients of the hedonic function, and therefore the implicit prices, are stable.

### **A promising approach**

The fact that clustering on the fitted values and residuals from the pooled model gives slightly better results between all clustering methods yields some insightful reflections. In effect, this approach is the only one that links the data clustering step to the estimation of the hedonic model, as residuals and predicted values carry information about the amount of bias in each observation was caused by estimating a single equation for all wine classes.

This brings us to formulate an approach that stands between modeling changing coefficients formally on the right-hand side and the data clustering approach. The general idea is to invert the order of the steps of the clustering approach and first obtain observation-specific estimates of the hedonic price function, and then use a clustering algorithm on the estimated coefficients. The final step is again estimating via OLS a relatively simple hedonic function specific to the identified classes.

Observation-specific estimates of the coefficients can be obtained via a nonparametric approach involving Kernel (or local) regression. It should be noticed that, just as in the clustering approach, the chosen Kernel function will calculate distances across observations to identify the neighborhood of data in which the hedonic function will be estimated, and then use the distances to weight assign more weight to observations closer to the point at which the function is evaluated, and less to the ones that are more distant. Thus, the process is still based on the intuitive idea that wines with a similar vector of attributes will belong to the same class and have similar implicit prices of the attributes, but this assumption is far less strong when imposed on a small neighborhood of the data, as it is in Kernel regression. If any out of sample is available, it is also possible to apply the weighting Kernel function only to the variables that are considered to be more influential in the product class differentiation process.

For this preliminary study we obtain observation specific-estimates by a local (first order) polynomial approximation, following the Lowess fit criterion proposed by Cleveland (1988). Lowess fit possesses several features that make it attractive for the task at hand: first, it is still based on (weighted) least squares, which is the estimator we use in the final step of cluster specific hedonic function; second, it uses a variable bandwidth that includes in the neighborhood a fixed (user-determined or data driven) percentage of observations that are closest to the point at which the function is evaluated; lastly, re-weighting procedures have been developed to yield local estimates of the hedonic function that are robust to the presence of outliers.

After the matrix of estimated parameters specific to each data point was obtained, the Ward algorithm (correlation matrix) was chosen to produce data cluster. Notice that in this case the ESS criterion being minimized becomes:

$$3.1. \quad ESS_i = \sum_{j=1}^N (\beta_j - \bar{\beta})'(\beta_j - \bar{\beta})$$

Again,  $i$  indexes the clusters ( $i=1, \dots, k$ ),  $\beta_j$  is vector of parameters associated with the  $j_{th}$  observation and  $\bar{\beta}$  is the vector of within cluster parameter means. This statistic is obviously related to the sample variance of the matrix of estimated parameters. In summary, the proposed methodology strives to find the partition of the data that constrains in the least possible way a linear model estimated for all the observations contained in a cluster.

Results of an explorative implementation of the methodology are reported in table 5. The table shows the price distribution within the clusters resulting from a single iteration (without robust re-weighting) of the Lowess fit for the testing sample. The size of the bandwidth was determined using the AICC<sub>1</sub> criterion proposed by Hurvic and Simonoff (1998). It can be noticed that the resulting clusters can easily be labeled as wine classes using the median price criterion mentioned earlier. Research on this method is ongoing as we write and results for the larger working sample and the resulting hedonic models are not yet available.

While the proposed methodology is computationally intensive, it is relatively simple to implement. It seems that between the inconclusive approach of data clustering and the complex parameterization of the hierarchical models, a methodology producing simple models in the presence of product classes would be a useful addition to the

literature. Off course, much more work is needed before anything can be concluded about the effectiveness of the approach, but these very preliminary results are encouraging.

**Table 5:** within cluster price distribution for the testing sample, local regression approach

Cluster	N	Mean	Median	Min	Max	Q1	Q3
Commercial	1382	17.39	16.35	3.39	77.25	12.36	20.60
Semi Premium	688	25.55	24.00	10.00	101.00	18.54	30.00
Premium	346	31.66	29.69	10.00	115.54	22.00	36.01
Ultra Premium	849	52.77	41.60	11.00	1014.00	30.90	60.00

## **References:**

- Allenby, G. and P. Rossi. 1999.** "Marketing Models of Consumer Heterogeneity"  
*Journal of Econometrics*, 89: 57-78.
- Angulo, A. M., Gil, J.M., Gracia, A., and Sanchez, M. 2000.** "Hedonic Prices for Spanish Red Quality Wine." *British Food Journal* 102(7):481-493.
- Bourassa, S.C., Hamelink, F., Hoesli, M., and MacGregor, B.D. 1999.** "Defining Housing Submarkets." *Journal of Housing Economics* 8:160-183.
- Bourassa, S.C., Hoesli, M., and Peng V.S. 2003.** "Do housing Submarkets Really Matter?" *Journal of Housing Economics* 12:12-28.
- Cleveland, W.S., Delvin, S.J., and Grosse, E. 1988.** "Regression by Local Fitting." *Journal of Econometrics* 37:87-114.
- Combris, P. and Lecocq S. and Visser, M. 1997.** "Estimation of a Hedonic Price Equation for Bourdeaux Wine: does quality matter?" *The Economic Journal* 107(441):390-403.
- Combris, P. and Lecocq S and Visser, M. 2000.** "Estimation of a Hedonic Price Equation for Burgundy Wine." *Applied Economics* 32:961-967.
- Costanigro, M., J. McCluskey and R. Mittelhammer. 2005.** "Segmenting the Wine Market Based on Price: Hedonic Regression When Different Prices Mean Different Products". Submitted to the *Journal of Agricultural Economics*.
- Dale-Johnson, D. 1982.** "An Alternative Approach to Housing Market Segmentation Using Hedonic Price Data". *Journal of Urban Economics* 11(3):311-332.
- Ernst and Young Entrepreneurs. 1999.** "Etude des Filières et des Stratégies de Développement des Pays Producteurs de Vins dans le Monde: Analyse de la

- Filiere Viticole Australienne”. *ONIVINS (Office National Interprofessionnel des Vins)*.
- Goodman, A.C., and T.G. Thibodau. 1998.** “Housing Market Segmentation”. *Journal of Housing Economics* 7(2):121-143.
- Greene, W. 2001.** “Fixed and Random Effects in Nonlinear Models”. *Working Paper EC-01-01*, Stern School of Business, Department of Economics.
- Hall, J. and Lockshin, L. and O’Mahony, G. B. 2001.** “Exploring the Links Between the Choice and Dining Occasion: Factors of Influence.” *International Journal of Wine Marketing* 13(1):36.
- Hurvich, C.M., and Simonoff, J.S. 1998.** “Smoothing Parameter Selection in Nonparametric Regression Using an Improved Akaike Information Criterion.” *Journal of the Royal Statistical Society B* 60:271-293.
- Landon, S. and Smith, C.E. 1997.** “The Use of Quality and Reputation Indicators by the Consumers: The Case of Bordeaux Wine.”, *Journal of Consumer Policy* 20:289-323.
- Oczkowski, E. 1994.** “Hedonic Wine Price Function for Australian Premium Table Wine.” *Australian Journal of Agricultural Economics*. 38, 93-110.
- Rosen, S. 1974.** “Hedonic Prices and Implicit Markets: Product differentiation in Pure Competition.” *Journal of Political Economy* 82:34-55.
- Schamel, G. and Anderson, K. 2003.** “Wine Quality and Varietal, Regional and Winery Reputations: Hedonic prices for Australia and New Zealand.” *The Economic Record* 79(246).

- Spawton, T. 1991.** “Marketing Planning for Wine.” *European Journal of Marketing* 25(3):2-47.
- Straszheim, M. 1974.** “Hedonic Estimation of Housing Market Prices: A Further Comment.” *Review of Economics and Statistics* 56(3):404-406.
- Watkins, K. 1999.** “Property Valuation and the Structure of Urban Housing Markets”. *Journal of Property Investment and Finance* 17(2):157-175.
- Wilhemsson, M. 2004.** “A Method to Derive Housing Sub-Markets and Reduce Spatial Dependency”. *Property Management* 22(3/4):276-288.