Asymmetric Price Effects of Competition[1]

Saul Lach
The Hebrew University and CEPR

José Luis Moraga-González
University of Groningen

March 18, 2009

[1]We are indebted to Pim Heijnen and Marco Haan for providing us with the gasoline price data and the list of gas stations operating in the Netherlands. Thanks also go to Corine Hoeben for giving us the data on municipal taxes in the Netherlands. David Genesove, Marielle Non, Dirk Stelder, Matthijs Wildenbeest and Shlomo Yitzhaki gave us useful comments. We also thank seminar participants at Ben Gurion University, CEMFI and at The Hebrew University. Yaron Aronshtan, Danny Bahar, Eli Berglas and Sarit Weisburd provided excellent research assistance. Moraga-González gratefully acknowledges financial support from the Marie Curie Excellence Grant MEXT-CT-2006-042471. E-mails: <saul.lach@huji.ac.il> and <j.l.moraga.gonzalez@rug.nl>.

Abstract

This paper examines how the distribution of gasoline prices (the minimum, median and maximum prices, as well as other quantiles) changes with the number of competitors in the market. Using data from the Netherlands we find that as competition increases, the distribution of prices spreads out: the low prices go down while the high prices go up, on average. As a result, competition has an asymmetric effect on prices. These findings, which are consistent with theoretical models where consumers differ in the information they have about prices, imply that consumers' gains from competition depend on their shopping behavior. In our data, all consumers, irrespective *of the number of prices* they are informed about, would benefit from an increase in the number of gas stations. The magnitude of the welfare gain, however, would be greater for those consumers that are aware of more prices. We conclude that an increase in the number of gas stations has a positive but unequal effect on the welfare of consumers in the Netherlands.

1 Introduction

Economists have dedicated a significant amount of effort to analyzing the relationship between the number of firms and prices. Standard oligopoly models assume consumers are perfectly informed about all prices in the market and predict that an increase in the number of firms will lower the equilibrium price. Alternative and more realistic models depart from the assumption that all consumers have the same information and describe equilibria characterized by non-degenerate price distributions.[1] In markets with price dispersion the question of what happens to "the" price when the number of firms changes is not even well defined. An increase in the number of firms usually affects the sellers' pricing strategies and this alters the whole distribution of equilibrium prices.

Empirical research of markets with price dispersion has usually proceeded by estimating the impact of competition on the mean and variance of prices.[2] In this paper we take a broader view and study how the distribution of prices changes with the number of competitors in the market. We examine how the minimum, median and maximum prices, as well as other quantiles of the price distribution, vary with the extent of competition as measured by the number of firms operating in a market. Specifically, we analyze the case of gasoline prices in the Netherlands.

We think this broader approach is important for at least two reasons. First, the class of theoretical models based on imperfect consumer information and search costs often predict that the effect of competition on "high" prices differs from its effect on "low" prices.[3] These models imply that the probability of observing relatively low *and* high prices increases with the number of firms operating in the market. Although this is a known theoretical result, to the best of our knowledge, it has not been verified empirically. This is surprising because, if empirically valid, this result has important welfare implications. Analyzing these implications constitutes the second motivation for

[1] See Baye et al. (2006) for a recent survey of models that rationalize price dispersion.

[2] See, for example, Borenstein and Rose (1994), Barron, Taylor and Umbeck (2004), Baye et al. (2004), and Lewis (2008).

[3] See, for example, Varian (1980), Stahl (1989) and Janssen and Moraga-González (2004). In these models some consumers know all the prices in the market while others only know one or two prices.

this paper. In markets exhibiting a single price in equilibrium, an increase in the number of firms reduces price and this unambiguously increases welfare for all consumers. When price dispersion is prevalent, different consumers may experience distinct welfare effects depending on how different parts of the price distribution respond to changes in the number of competitors. If, as these models predict, the frequency of low *and* high prices increases with competition then whether consumers are successful in paying the lower prices depends on their shopping behavior. Increased competition is likely to favor more those consumers sampling or observing several prices because they may end up paying one of the lower prices. By the same token, increased competition may even hurt consumers that observe very few prices (e.g., only one price) because they may end up paying one of the higher prices. Theoretically, price changes originating from an increase in the number of firms can result in welfare gains for some consumers and at the same time in welfare losses for others. Analyzing the effects of entry-promoting policies just on the mean and dispersion of prices cannot capture these distinct welfare effects.

We use daily Euro 95 price data posted by about 3100 gas stations in the Netherlands during May 2006.[4] For a given gas station, the relevant market is defined as the municipality where the gas station is located. For each of such 423 markets, we compute the minimum, median and maximum price, as well as other quantiles of the price distribution. We then regress these statistics on the number of gas stations in the market as well as on municipality characteristics to control for common determinants of prices and the number of stations. We also use population size and local taxes as instruments for the (endogenous) number of stations.

The empirical findings suggest that as competition increases the distribution of prices spreads out; therefore competition has *asymmetric effects on prices*. Specifically, as the number of gas stations in a market increases, the low prices decrease while the high prices increase, on average. Adding 4 additional gas stations to a single-station market would, on average, lower the minimum price of a liter of Euro 95 by 0.93 cents and increase the maximum price by 0.83 cents. These are small changes relative to the

[4]The results also hold for Diesel (see Section 5.1).

mean price of 142 cents, but these changes are quantitatively significant relative to the dispersion in (residual) prices which is about 1 cent. This characterization of the effect of competition on prices accords in part with the theoretical predictions of models where some consumers have imperfect information about prices and observe (perhaps through search) different number of prices.

In addition, we estimate the gains from increased competition to consumers observing different numbers of prices. In our data, all types of consumers benefit from an increase in the number of stations. The magnitude of the welfare improvement due to price changes depends, however, on their shopping behavior and is larger for those consumers that observe more prices. The decline in the expected price paid by consumers that observe 4 or 5 prices is about twice as large as that for consumers that observe only 2 prices.

We believe the message of this paper goes beyond the present application to the gasoline market in the Netherlands. Since imperfect price information is prevalent in many markets (telecommunications, health, gas, electricity, etc.), the price effects of competition-enhancing policies (industry deregulation, trade liberalization, etc.) might not be as straightforward as those implied by standard models. Moreover, since increased competition can potentially have unequal effects among consumers, distributional issues become a central part of the welfare assessment of these policies. This advocates the importance of taking a broader view where the interaction between competition and consumer policy is taken into consideration (Armstrong, 2008; Waterson, 2003).

In the next section we present a Varian-style model of the distribution of prices in an oligopolistic market where consumers differ in the amount of prices they are exposed to. The model delivers implications on the effect of the number of firms on the price distribution (minimum, maximum and mean price, as well as other quantiles), and on the price paid by various consumer types. The model also makes clear that increased competition has an effect on prices only when it increases the amount of information consumers have. Section 3 describes the gasoline price data for the Netherlands and explains how markets are defined for the empirical analysis. We present evidence that gas stations in our data appear to be using mixed pricing strategies as implied by Varian-style

models. We also show preliminary graphical evidence on the relationship between the number of gas stations and the minimum and maximum price. Our empirical strategy is outlined in Section 4, while the empirical results are presented in Section 5. An empirical assessment of the welfare implications of increased competition is presented in Section 6. Conclusions close the paper.

2 A model of the distribution of prices

The market for gasoline is a good example of a homogenous good market where price dispersion is observed.[5] Many consumers are informed about a few prices only, and this gives monopoly power to the gas stations. In many instances, consumers run out of fuel and have no option but to fill their gas tanks at the first gas station they encounter and this also gives market power to the gas stations. Prices change quite frequently and it is not trivial to tell which gas station is the cheapest in a given market. The model we consider below, an extension of Varian's (1980) model of sales, has these characteristics.

Suppose we have a market where there are $N \geq 2$ identical firms that compete in prices to sell a homogeneous good to a large number L of consumers. We assume firms' unit selling costs, c, are the same across all firms.[6] Each consumer wishes to purchase at most a single unit of the good (e.g., a full tank). The maximum willingness to pay for the good is common across consumers and is denoted by $v > 0$. The entire population of consumers L can be divided into various groups or types, each group consisting of all the consumers with similar exposure to price information. In particular, we assume that a fraction μ_s of the consumers is informed about s prices in the market, with $s = 1, 2, ..., S$. The rationale behind this assumption is that the typical consumer is exposed to a number of prices that depends on the number of gas stations he/she observes while driving to work.

[5] Price dispersion in gasoline markets has been widely documented. Recent papers on this topic are, for example, Barron, Taylor and Umbeck (2004), Chandra and Tappata (2008), Hosken et al. (2008), and Lewis (2008).

[6] We therefore abstract from cost differences across firms as an explanation for price dispersion in gasoline markets. In the empirical part, however, we control for station-specific unobserved effects.

We shall assume that $S \leq N$. In small markets where there are up to, say, 6-8 gas stations it is reasonable to expect that S be equal to N. However, in cities where there is a very large number of gas stations (Rotterdam, for example, has 80 gas stations), it is unrealistic to assume the existence of consumers that are aware of all prices in the market as done, for example, by Varian (1980). In those markets S will typically be much smaller than N. As it will become clear later, it is S and not N that determines the extent of competition in the market.

Moreover, we believe that search in this market is "passive" in the sense that consumers do not deliberately drive to various gas stations to observe their prices. Nevertheless, we will refer to the different consumer types as consumers exhibiting different "shopping" behavior.[7] In addition, as we will see later, this assumption serves to produce empirically plausible price densities. It will be convenient to denote the fraction of consumers observing exactly S prices as γ; then by construction $\gamma = 1 - \sum_{s=1}^{S-1} \mu_s$.

Firms play a simultaneous-moves Bertrand game. An individual firm i chooses a price p_i taking the prices of the rival firms as given. To rule out pure-strategy equilibria, we shall assume $1 > \mu_1 > 0$ (as in Varian, 1980).[8] The intuition is as follows. Consider the position of a firm i and suppose all its rivals were charging a price $\widetilde{p}$, with $c \leq \widetilde{p} \leq v$. There are two forces that affect the price-setting decision of firm i. First, there is a desire to *steal business* from competitors and this pushes this firm to undercut the rivals' prices. This desire arises because there exist consumers who are exposed to various prices and choose the cheapest gas station to tank (i.e., $\mu_s > 0$ for at least one s, $2 \leq s \leq S$). Second, the possibility of *extracting surplus* from consumers who do not compare prices prompts firm i to set higher prices than the rivals. This desire arises because there exist consumers (in particular a fraction $\mu_1/N > 0$) who have no other option but to tank at firm i. It is easy to see that either of these deviations destabilizes any such price $\widetilde{p}$.

[7] We therefore do not model the consumer's decision of how many prices to observe. For models of this kind where there are two types of consumers see Stahl (1989) and Janssen and Moraga-González (2004).

[8] If $\mu_1 = 1$, then $p_i = v$ for all i is a pure-strategy equilibrium. If $\mu_1 = 0$, then $p_i = c$ for all i is a pure-strategy equilibrium.

Therefore a single price cannot accommodate these two incentives.

Denote the mixed pricing strategy of a firm i by a distribution of prices F_i. We shall only study symmetric equilibria, i.e., equilibria where $F_i = F$ for all $i = 1, 2, \ldots, N$.[9] To calculate the expected profit to firm i from charging a price p when its rivals choose a random pricing strategy according to the cumulative distribution F, we first consider the chance that firm i sells to a consumer of type s, i.e., to a consumer that observes s prices in the market. The chance that such a consumer observes the price of firm i is s/N and, conditional on this, the probability that firm i sells to this consumer at price p is $(1 - F(p))^{s-1}$. Therefore, the profits to firm i from all types of consumers is

$$\Pi_i(p; F) = L(p - c) \left[\sum_{s=1}^{S-1} \frac{s\mu_s}{N} (1 - F(p))^{s-1} + \frac{S\gamma}{N} (1 - F(p))^{S-1} \right] \tag{1}$$

In equilibrium, a firm must be indifferent between charging any price in the support of F and charging the upper bound $\overline{p}$. Therefore, any price in the support of F must satisfy $\Pi_i(p; F) = \Pi_i(\overline{p}; F)$. Since $\Pi_i(\overline{p}; F)$ is monotonically increasing in $\overline{p}$, it must be the case that $\overline{p} = v$. As a result, equilibrium pricing requires

$$(p - c) \left[\sum_{s=1}^{S-1} s\mu_s (1 - F(p))^{s-1} + S\gamma (1 - F(p))^{S-1} \right] = (v - c)\mu_1 \tag{2}$$

This equation cannot be solved explicitly for F, except in special cases. However, the existence of a equilibrium price distribution F can be easily proven. This symmetric equilibrium is unique.[10]

We are interested in how the pricing strategy of the firms changes with the number of firms N. For this, we need to distinguish markets with many firms ($N > S$) from markets with a few firms ($N = S$). In the former case, an increase in the number of firms would have no effect on pricing behavior. This is because consumers would observe at most S prices and so increasing the number of stations would not affect consumers'

[9] It is easy to see that the support of F must be a convex set and that F cannot have atoms.

[10] To prove that F exists, rewrite equation (2) as $\sum_{s=1}^{S-1} s\mu_s (1-F(p))^{s-1} + \gamma S (1-F(p))^{S-1} = \frac{\mu_1 (v-c)}{p-c}$. Since $F \in [0, 1]$, the LHS of this equation is positive and decreases in F; by contrast, the RHS is a positive constant. At $F = 0$, the LHS equals $\sum_{s=1}^{S-1} s\mu_s + \gamma S > 0$, while at $F = 1$ it takes on value μ_1. As a result, for every $p \in (\underline{p}, v)$, there is a unique solution to equation (2) satisfying $F \in [0, 1]$ and F increases in p.

information – equation (2) does not depend on N. Of course, firms' profitability would be affected by N but not firms' pricing strategies.

Consider the more interesting case of a market with few gas stations, i.e., where $S = N$. In this case there are always consumers who are informed about all the prices in the market. Let F_N be the equilibrium price distribution that solves equation (2) (we index F by N to indicate that the pricing strategy does indeed depend on the number of competitors in the market). The lowest price charged in the market, denoted $\underline{p}_N$, can be found by setting $F(\underline{p}_N) = 0$ in equation (2) and solving for

$$\underline{p}_N = c + \frac{(v-c)\mu_1}{\sum_{s=1}^{S-1} s\mu_s + \gamma N} \tag{3}$$

Since our equilibrium satisfies equation (2), the comparative statics of the price distribution with respect to the number of firms can be obtained by applying the implicit function theorem to that equation (treating N as a continuous variable for simplicity)

$$\frac{dF_N(p)}{dN} = \frac{\gamma(1-F_N(p))^{N-1}\left[1 + N\ln[1-F_N(p)]\right]}{\sum_{s=1}^{S-1} s(s-1)\mu_s(1-F_N(p))^{s-2} + \gamma N(N-1)(1-F_N(p))^{N-2}} \tag{4}$$

The denominator of this expression is clearly positive. Therefore

$$sign\left[\frac{dF_N(p)}{dN}\right] = sign\left[1 + N\ln[1-F_N(p)]\right] \tag{5}$$

We now make some observations on the expression $1 + N\ln[1-F_N(p)]$. This expression is monotonically decreasing in p. When $p \to \underline{p}_N$, we have $F_N(p) \to 0$ and the sign of (4) is positive. By contrast, when $p \to v$, we have $F_N(p) \to 1$ so the sign of (4) is negative. Therefore there exists a unique price $\widehat{p}_N$, such that $dF_N(p)/dN \geq 0$ for all $p < \widehat{p}_N$ and $dF_N(p)/dN \leq 0$ for all $p > \widehat{p}_N$. In fact, $\widehat{p}_N$ satisfies $F(\widehat{p}_N) = 1 - \exp[-1/N]$.

The implication of this result is that an increase in the number of competitors results in an increase in the probability with which low and high prices are charged and, therefore, in a decrease in the probability with which intermediate prices are charged. This can clearly be seen in Figure 1 where we have plotted the equilibrium price distributions for markets with 2 firms (thick solid curve), 4 firms (thin solid curve) and 6 firms (dashed curve), for given parameters.[11]

[11] The parameters are $\mu_1 = 0.5$, $\mu_2 = 0.4$ and $\gamma = 0.1$; the rest of the μ's and the marginal cost c are

In order to understand the intuition behind these effects, recall that the distribution of prices of a firm is chosen to maximize the profits accruing from the various groups of consumers, given other firms' strategies. These profits, as shown is equation (2), are constant at all prices chosen with positive probability. Note also that the elasticity of the *expected* demand of the consumers observing $s < N$ prices is equal to $p(s-1)f_N(p)/(1-F_N(p))$.[12] Thus, keeping rivals' strategies fixed, this elasticity is independent of N. By contrast, the elasticity of the expected demand of consumers observing N prices is $p(N-1)f_N(p)/(1-F_N(p))$, which increases in N. Therefore, if the number of firms increases, and keeping the rivals' strategies fixed, only the elasticity of the expected demand of the fully informed consumers changes. Consider a firm contemplating how to change its strategy in response to entry. This firm knows that the expected demand from the fully informed consumers becomes more elastic as N increases; it therefore has an incentive to offer even lower prices when N increases. Thus, in this model, the competitive effect of an increase in N is to prompt firms to offer "lowest ever" prices by shifting some probability mass from initially low prices to even lower prices (prices that had zero density before the increase in N). This effect is clearly visible in Figure 1 where it is seen that the lowest price in the support of the price distribution, $\underline{p}_N$, decreases with N.

At the same time, it becomes disproportionately more difficult to sell to fully-informed consumers because there are more competitors in the market. Thus, firms tend to "give up" on selling to the fully-informed consumers in favor of selling to the less-informed consumers. The latter's lack of full information on prices allows firms to increase the frequency at which they charge high prices..This effect is also visible in Figure 1 where it is seen that the slope of F_N near the upper bound v increases with N.

In sum, a change in N induces non-trivial changes in the equilibrium price distribu-

---

set equal to zero, while $v = 1$. Therefore, when $N = 2$, half of the consumers observe one price, and the other half $(\mu_2 + \gamma)$ compare two prices. In the $N = 4$ case, we have again that half of the consumers observe one price, 40 percent observe two prices and 10 percent of the consumers observe four prices.

[12]Every consumer demands at most a single unit so individual demands are inelastic. However, given the strategies of the rival firms, the expected number of units sold by a firm increases as the firm reduces its price.

tion since it comprises changes in the support of the price distribution as well as changes in the price frequencies. Overall, however, pricing strategies become more extreme and, as we state next, the mean price increases with N (the proof is in the Appendix).

Proposition 1 *Consider the case of a market where the number of firms is relatively small, i.e., $S = N$. Then the following results hold:*

(A) Let p_N be distributed according to the equilibrium price cdf $F_N(p)$; likewise, let p_{N+1} follow the distribution $F_{N+1}(p)$. Then $E[p_{N+1}] > E[p_N]$.

(B) Let $y_N = \min\{p_1, p_2, ..., p_N\}$ where where $p_1, p_2, ..., p_N$ are i.i.d. random variables drawn from $F_N(p)$; likewise, let $y_{N+1} = \min\{p_1, p_2, ..., p_{N+1}\}$ where $p_1, p_2, ..., p_{N+1}$ are i.i.d. random variables drawn from $F_{N+1}(p)$. Then, there exists a number $\overline{\gamma}(N)$ such that $E[y_N] > E[y_{N+1}]$ for all $\gamma < \overline{\gamma}(N)$.

(C) Let $z_N = \max\{p_1, p_2, ..., p_N\}$ where where $p_1, p_2, ..., p_N$ are i.i.d. random variables drawn from $F_N(p)$; likewise, let $z_{N+1} = \max\{p_1, p_2, ..., p_{N+1}\}$ where $p_1, p_2, ..., p_{N+1}$ are i.i.d. random variables drawn from $F_{N+1}(p)$. Then, there exists a number $\widehat{\gamma}(N)$ such that $E[z_{N+1}] > E[z_N]$ for all $\gamma > \widehat{\gamma}(N)$.

If, instead, the market has a relatively large number of firms, i.e., $S < N$, then an increase in the number of firms has no effect on the price strategies of the firms.

It is important to remark that the conditions in Proposition 1are sufficient but not necessary. A large number of simulations indicate that it is always the case that the expected minimum price decreases with N while the expected maximum price increases with N. In other words, we find that for all sensible values of γ Proposition 1 holds unconditionally.

The effect of the number of competitors on the price distribution constitute the first set of implications of the model that we will test using gasoline prices in the Netherlands.

A comment is in order here. The results in Proposition 1 have been derived assuming consumer behavior is fixed. By "consumer behavior" we mean that the distribution of the partially-informed consumers, that is $\mu = (\mu_1, \mu_2, ..., \mu_{S-1})$ does not change with N. This implies that when we move from a market with N firms to a market with $N+1$

firms the only thing that changes is that the fully-informed consumers move from observing N prices to observing $N+1$ prices.[13] It is precisely because some consumers observe all the prices in the market that an increase in the number of firms has an effect on prices. If the number of firms were to increase but no consumers observe all the $N+1$ prices, then increased competition would have no effect on prices. Thus, increased competition has an effect on prices only when it increases the amount of information consumers have.

A second, but not less important, issue is the evaluation of the welfare effects of increased competition implied by the model. In other words, what happens to the prices actually paid by the different consumers when N increases?

Proposition 1 above already provides a partial answer to this question. Consumers that observe only one price pay, on average, the mean price $E[p]$. This price increases with N so these consumers are, on average, worse-off.[14] Consumers that observe all N prices pay, on average, $E[y_N]$ which turns out to decrease with N. Other consumers, those observing $1 < s \leq S$ prices, pay $y_s = Min\{p_1, p_2, ..., p_s\}$ each time they purchase, where $p_1, p_2, ..., p_s$ are i.i.d. random variables drawn from F_N. On average, they pay $E[y_s] = E[Min\{p_1, p_2, ..., p_s\}]$, where the expectation is taken over the distribution F_N. It turns out that $E[y_s]$ can increase or decrease with N depending on parameters. It therefore becomes an empirical matter whether these consumers pay a lower or higher price after the number of competitors increases.

Figure 2 illustrates the relationship between the expected prices paid and N using the same parameters as in Figure 1. In this simulation we set $S = 7$ and it becomes apparent that prices do not change when N increases beyond 7. Inspection of the graph reveals that consumers who are typically informed about two prices pay on average

[13] Admittedly this is a simplification. In an endogenous search model, an increase in the number of firms will in general lead to an overall change in the shares of consumers observing s prices. As shown in Janssen and Moraga-González (2004) this may cause a price statistic to behave non-monotonically with respect to the number of firms.

[14] This tendency to raise prices on average was first shown by Stahl (1989) in a model with two consumer types (shoppers, who observe all the prices in the market, and non-shoppers, who just observe one). In his model, search is endogenous and the price distribution converges to a distribution degenerated at the monopoly price as the number of firms operating in the market goes to infinity.

$E[y_2] = E[\min\{p_1, p_2\}]$, which, for this configuration of parameters, increases slightly with the number of firms. The graph also reveals that the mean price also increases while the average price paid by the fully-informed consumers goes down.

The welfare effects of an increase in the number of firms are therefore complex. While some consumers benefit others may lose. This ambiguous effect is a direct result of the way the equilibrium price distributions change with N (i.e., the distributions with N and $N+1$ firms cannot be ranked according to first-order stochastic dominance). It is therefore difficult to evaluate the change in aggregate welfare without making assumptions on the importance of the gains relative to the losses from increased competition.

2.1 Price densities

The equilibrium price densities corresponding to the cumulative density functions in Figure 1 are depicted in Figure 3. It can be seen that the model allows for bell-shaped density functions. This is a desirable implication of the model because bell-shaped densities are a typical feature of real-world price data. In particular, Figure 4 shows that the density of gas prices in the Netherlands is bell-shaped. Other studies have also found bell-shaped price density functions for various products (e.g., Lach, 2002, Hosken et al., 2008).

Varian's (1980) model corresponds to the case where there are only two types of consumers: fully-informed consumers γ and uninformed consumers $\mu_1 = 1 - \gamma$.[15] In this case, the price density is either decreasing or U-shaped (see e.g. Figure 2 in Varian (1980)). Thus, because empirical price densities are usually bell-shaped, the simple Varian model is inconsistent with the data. Assuming additional consumer heterogeneity in the form of additional consumer types, besides being a more realistic assumption,

[15] Under these assumptions we can solve explicitly for $F_N(p)$, namely,

$$F_N(p) = 1 - \left(\frac{(1-\gamma)(v-p)}{N\gamma\,(p-c)} \right)^{\frac{1}{N-1}}$$

with support

$$c + \frac{(1-\gamma)(v-c)}{(1-\gamma)+N\gamma} \leq p \leq v$$

allows for bell-shaped price density distributions to arise in equilibrium.

2.2 Exogenous changes in N

In equilibrium, the number of competitors N is likely to be determined in part by factors that also affect prices, such as income (which determines v) and the distribution of consumer types. For example, if we compare two markets with different willingness to pay, v, then the market with higher v is likely to have both a higher N and higher prices thereby generating a positive relationship between prices and the number of stations. This relationship, however, would not be a causal one if the effect of v on both N and prices is not accounted for, which is likely to be the case since v is inherently unobserved.

If we want to know the causal effect of a change in N on the distribution of prices, we need to ensure that changes in N are not accompanied by changes in other determinants of $F_N(p)$. The existence of such exogenous variation in N is crucial for interpreting the effect of a change in the number of firms on prices. In this context, it is important to recall that the effect of a change in N works through the additional information that the fully-informed consumers have when they become aware of more prices. That is, what we refer to as the causal effect of N can only arise when it induces a change in the information consumers have.

Economic theory is very clear on the determinants of the number of firms N. Namely, in a long-run market equilibrium with free entry, the number of stations is determined by a zero-profit condition, where profits are net of entry costs. Let $\mu = (\mu_1, \ldots, \mu_{S-1})$. From equation (1) we see that the profits of a firm are determined by c, μ, N, v and L. But, from (2), we have that the equilibrium price distribution F_N is itself determined by c, μ, N and v. It is therefore only when N changes because of changes in L (and/or, as seen below, in entry costs) that the variation in N is exogenous to prices in the sense that nothing else affecting the price distribution changes when such a change in N occurs.[16]

More formally, if we denote expected profits per consumer by π and express π as

[16]The assumption of constant returns to scale in selling gasoline is important here. Otherwise retail costs would depend on the number of consumers L.

a function of N and the exogenous variables in the model, the zero profit condition is

$$L\pi(N, c, v, \mu) - E = 0 \tag{6}$$

where E denotes entry costs.

Thus, L and E are the natural instruments for N since they help in determining the number of firms but do not affect prices. Changes in L or in E (or in both), create exogenous variation in N which will allow us to estimate its causal effect on prices.

3 The price data, mixed strategies and preliminary evidence

We have daily prices for Euro 95 gasoline in a large sample of gas stations in the Netherlands.[17] The price data were obtained from Athlon Car Lease Nederland B.V., the largest private car leasing company in the Netherlands with over 129.000 cars as of the end of 2008 (www.athloncarlease.com). The typical contract between Athlon and its lessees stipulates that Athlon pays for the gasoline consumed (up to a limit) as well as for car maintenance, insurance, etc. In order to do this, the lessees submit their gas receipts to Athlon and it is from these receipts that the gasoline prices are retrieved. These prices are therefore actual prices paid by drivers at the pump. It should be remarked that all gas stations are self-service, that there are no discount prices for Athlon's lessees and that lessees have no incentives to search for and fill up at the station offering the lowest price. This last point is important because it allows us to view the sample of prices as randomly drawn.

Prices were obtained from 3,300 gas stations for the period May 5–26 2006, except for May 10 and May 17, for a total of 20 days. 217 stations located in highways were deleted from the sample because the model guiding our empirical analysis is not relevant to these stations, leaving us with price data from 3,083 gas stations. Because the price information arrives directly from the lessees, not all stations are sampled every day,

[17] We also have data for Diesel. We use these data in Section 5.1 to show that our results also hold for a different gas product.

which results in an unbalanced panel data of gas stations.[18] There are 32,348 station-day observations on Euro 95 prices.

The location of each station is given by a 4-digit zip code. It is clear that identifying a market by the area in a single 4-digit zip code is too narrow since it is likely that many people cross several 4-digit zip code areas while commuting to work and observe prices in different zip codes. We therefore expand the geographic coverage and define a market to be the area comprised by a municipality. A municipality is a group of 1 or more 4-digit zip code areas sharing common borders. There are 440 municipalities in the Netherlands for which we have gasoline price data. About 13 percent of the municipalities cover exactly one zip code area, while 54 percent are comprised of up to three zip code areas.[19] The majority of the municipalities are quite small in terms of population: 55 percent have less than 25,000 inhabitants, and the population in 91 percent of the 440 markets is under 75,000.

This definition of the market is clearly not perfect as it ignores stations that may be geographically close (or in the way to work) but located in different municipalities. This may not constitute much of a problem in our model where the measure of competition is S, the maximal number of prices consumers in the market observe. What is important is that every gas station in a given municipality factors the same number of competitors S into its pricing strategy– it is not necessary for these S stations to be located in the given municipality. In the empirical work we proxy S by N, the actual number of gas stations in the municipality, to estimate how changes in N affect the price distribution. In any case, we examine the robustness of our findings to the inclusion of the number of gas stations in neighboring municipalities in Section 5.1.

More importantly, another reason for choosing to work at the municipality level is that we have economic, geographic and demographic data for almost each municipality,

[18] We have one price of Euro 95 per station per day. The number of days or, equivalently, the number of price quotations per gas station in the sample ranges from 1 to 17 days with an average of 10.5 days and a median of 12 days.

[19] On average, a municipality covers 4.4 4-digit zip code areas. The largest municipalities (in terms of number of 4-digit zip codes covered) are Amsterdam, Rotterdam, and 's-Gravenhage with, respectively, 44, 41 and 28 4-digit zip code areas.

while not all this information is available at the zip code level.[20] This is very convenient for our purposes since we will be able to control for common determinants of the number of stations and prices.

We obtained a list of *all* the gas stations and their addresses operating in the Netherlands in August 2007. These stations were assigned to municipalities according to their addresses. This allows us to know the number N_m of gas stations operating in a market m. We do not have price data on all N_m stations and therefore N_m is at least as large as the number of stations with price data in the sample.[21] The mean number of stations by market is 8.2, respectively, and there is a lot of variation across markets – the standard deviation is almost as large as the mean, 7.6 stations. This variation is better seen in Table 1 where the distribution of the number of stations per market (municipality) is tabulated. N_m ranges between 1 to 80 (Amsterdam has 59 stations and Rotterdam has 80). Sixty percent of the markets have 7 or less stations.

As an illustrative device, the left panel in Figure 4 displays a kernel estimate of the density function of prices. The average price of Euro 95 gas in our sample is 142.04 cents and the standard deviation is 3.37 cents. The lowest price is 119 cents while the highest price in the sample is 167.[22] Not surprisingly, there is dispersion in gasoline prices but, as evidenced by the coefficient of variation, it is not very large. However, the daily variation in the total cost of filling-up a 50 liter tank – the difference in cost between buying at the highest-priced and lowest-priced station in a given day – is between 8.5 and 24 euros which is not a trivial amount.[23]

Because price differentials among stations are likely to be driven by time-invariant

[20] Data obtained from Statistics Netherlands (www.cbs.nl).

[21] This is because we observe gas prices only from Athlon's lessees who do not patronize *all* the gas stations.

[22] The 119 price is an outlier; the second lowest price is 129 cents. However, we do no think this is a typo since the very same gas station is also charging a very low price for Diesel on a *different* day (78 cents when the average is 108 cents).

[23] On May 11, the maximum and minimum price were 1.49 and 1.32 euros per liter. This 17 cent difference translates into a 8.5 euros saving for a full 50 liter tank. On May 14, the maximum and minimum price were 1.19 and 1.67 euros per liter.

factors (e.g., brand, location, availability of a convenience store, additional services, etc.), it is problematic to compare prices of different gas stations, even within the same market. The same is true when comparing gas prices in different days. We therefore remove day and station-specific effects from actual (raw) prices to obtain a residual price which is more comparable across days and stations.[24] These residual prices are obtained by regressing prices on station-specific dummies and on a cubic trend, separately for each municipality.[25] Residual prices are therefore detrended prices net of station-level effects. The mean residual price for each station (and for each municipality) is then zero. The implicit assumption here is that station and day effects affect *only* the mean price charged by a gas station. By removing these effects, we "homogenize" stations within markets so that we can treat residual prices in a market as coming from the same distribution of prices F_N. Of course, the distribution of residual prices varies across markets due to differences in consumer shopping behavior and in the number of firms, as well as in v and c.[26]

We will use the residual prices in our empirical analysis; their distribution is plotted in the right panel of Figure 4. As expected, the standard deviation in residual prices, 1.07, is lower than that in the raw data. Nevertheless, as can be seen in the graph, residual prices still exhibit considerable variation.[27]

[24] As done, for example, by Lach (2002), Hosken et al. (2008), and Lewis (2008). A similar approach is taken when estimating auction data in order to generate "homogenized bids" or "normalized bids" which are comparable across auctions. See, for example, Haile, Hong, and Shum (2003) and Hu and Shum (2008).

[25] We do not use day dummies because in 5.4 percent of the observations there is only one station per day per municipality. Moreover, in one municipality (Reiderland) we only have one observation and therefore residual prices cannot be computed, leaving us with residual prices in 439 municipalities.

[26] Note, however, that the theoretical model in Section 2 implies that proportional changes in v and c affect only the location of F_N. In any case, removing store effects also removes the effect of market-level factors affecting the location of F_N.

[27] The longer left tail of the distribution is due to the outlier price mentioned in footnote 21. Removing this station makes the density much more symmetric and lowers the standard deviation to 1.06.

3.1 Evidence on mixed strategies

Before we examine the relationship between the number of stations and the distribution of prices we present empirical evidence on the use of mixed strategies by gas stations. This issue has been examined by Lach (2002), Wildenbeest (2008) and Hosken et al. (2008), for various products. All these studies found evidence in support of the use of mixed pricing strategies in different product markets. The Hosken et al. (2008) study is most relevant for our purposes because it examines the pricing behavior of 272 gas stations in the suburbs of Washington DC.

In this section we check whether gas stations vary their relative position in the cross-sectional distribution of prices over time, as implied by the use of mixed strategies. Simply put, the use of mixed strategies implies that we should not observe gas stations always selling at high prices or always selling at low prices.

We observe the residual price posted by gas station i on day t and we locate this residual price within the price distribution observed in the gas station's market for day t. We can then track the relative position of the station's residual price over time.[28] There are a number of ways of doing this.

We start by computing the number of days that a gas station was in the q^{th} quartile of the cross-sectional price distribution. We denote this statistic by T_q, $q = 1, 2, 3, 4$. T_q is expressed as a percentage of the total number of days a station appears in the sample.[29] For example, if the station was never in the first quartile of the distribution then $T_1 = 0$, whereas if the station was always in the first quartile then $T_1 = 1$. Clearly, for each gas station, $T_1 + T_2 + T_3 + T_4 = 1$. Figure 5 plots the histograms of $T_1 - T_4$.

If many stations always remain in the same quartile of the (residual) price distribution we should observe a large number of firms with $T_q = 1$. Figure 5 indicates that this is not the case. 2.15 percent of the stations were always in the first quartile and,

[28] Recall that the time horizon is the 20 days in May 2006 but no station appears in the sample for more than 17 days.

[29] Note that the cross-sectional distribution in day t is defined for the stations which quoted prices in day t. Therefore the number of stations, and their identity, may change from day to day. The statistics were computed in all markets and days where the number of stations was at least 4.

for higher quartiles, this percentage is even lower. The low number of stations always selling in the same quartile of the price distribution is consistent with the use of mixed strategies.[30]

We also observe in the top-left graph of Figure 5 that 14 percent of the stations were never in the first quartile of the price distribution ($T_1 = 0$). This means that about 84 percent ($100-14-2.15$) of the stations were part (but not all) of their time in the first quartile of the distribution and the remaining time in other quartiles. Similarly, 77, 81 and 75 percent of the stations were part (but not all) of their time in the second, third and fourth quartile of the distribution, respectively, and the remaining time in other quartiles.[31] Although this is evidence that a sizable number of stations moves around the cross-sectional price distribution, the histograms of $T_1 - T_4$ do not reveal how long a particular gas station stayed in each quartile of the price distribution. We examine this in Figure 6.

Figure 6 graphs, for 50 randomly selected gas stations, the percentage of days each of these stations was in the first, second, third and fourth quarter of the cross-sectional distribution of prices.[32] The changing bar colors indicate that only two stations remained in the same quartile during all the days they appear in the sample (stations number 2 and 9).[33]

These figures still do not reveal how gas stations "travel" across the quartiles of the price distribution over time, i.e., the extent of intra-distribution dynamics. The transition process from one cross-sectional distribution to another can be modelled by assuming that this transition is done in a Markovian fashion through a 4×4 transition

[30] When using the actual prices, 20 percent of the stations always charge prices in the first quartile of the distribution ($T_1 = 1$), 1.7 percent always in the second quartile ($T_2 = 1$), 3.4 percent always in the third quartile ($T_3 = 1$), and 7 percent always in the fourth quartile ($T_4 = 1$). These figures are higher because actual prices reflect store-specific factors (e.g., location) that are fixed over time.

[31] The corresponding percentages for the actual price data are 42, 54, 47 and 33 percent.

[32] We plot only 50 stations that were randomly sampled from the 2472 gas stations appearing in markets and days where the number of stations was at least 4. Plotting all the stations generates graphs that are too cluttered to be readable.

[33] Using the actual rpice data, 62 percent of the stations (31 stations) do move between 2 or more quartiles of the price distribution.

matrix whose $(i,j)^{th}$ entry gives the probability that a gas station in the i^{th} quartile in day t moves to the j^{th} quartile in day $t' > t$. Consistent estimates of these probabilities are the sample proportions of stations moving from one quartile to another. Assuming a time-invariant transition matrix, the estimated transition matrices for each day t can be averaged to produce a single (estimated) transition matrix.

Table 2 presents estimates of 1–week ($t' = t + 7$) transition probabilities.[34] Examination of the transition matrix gives a good idea on the extent of intra-distribution mobility. If stations keep their positions over time – lack of mobility – then the matrix should have "large" diagonal entries. If there is a lot of mobility across the quartiles of the distribution this would be reflected in "large" off-diagonal probabilities. The probability of remaining in the first quartile is 33 percent, which means that the probability that a low-price station will be selling at a higher price a week ahead is 67 percent.[35] Overall, the diagonal entries do not appear to be large relative to the off-diagonal terms. This is indicative of significant intra-distribution dynamics. Similar conclusions were reached by Hosken et al. (2008) for gasoline prices in the U.S., by Lach (2002) for other products in Israel, and by Wildenbeest (2008) for groceries in the UK.

3.2 Preliminary evidence on prices and N

One of our goals in this paper is to study the relationship between the number of gas stations (N) and the distribution of prices in a market. In particular, the model presented in Section 2 implies that the minimum price should decline with N, whereas the mean and maximum price should increase.

We now present graphical evidence on the relationship between the minimum and maximum prices and N, while in the next sections we present the econometric evidence. For each market, we compute the minimum and maximum residual price observed in all stations and over all days. Recall that residual prices are detrended so that we can pool

[34] The entries are weighted averages of the estimated transition probabilities for each day with weights equal to the proportion of observations in each cell.

[35] Using the actual price data we obtain probabilities of remaining in the same quartile a week ahead equal to .76, .47, .51 and .69 for the first, second, third and fourth quartile, respectively.

observations from different days. However, since residual prices add up to zero for each station and market by construction, their mean price by market is identically zero and, therefore, we cannot study how the mean price varies with N.

In Figure 7 we plot the minimum and maximum residual price against N. Each point represents a market. The plot suggests that the minimum price decreases with the number of stations in the market, while the maximum price increases with N. To get a quantitative feeling for the magnitude of these effects, we regressed the minimum and maximum residual price on $\ln(N)$ over all the municipalities. The estimated slope parameters are -0.51 and 0.37 with robust standard errors 0.052 and 0.034. These are significant effects. The semi-log specification implies that the marginal effect of N decreases with N, which seems appropriate for these data. The predicted values of these regressions are plotted by the solid lines in Figure 7.

In Table 3 we tabulate the averages of the minimum and maximum residual prices (the points in Figure 7) over all markets having the same N, as well as the average of other quantiles in the residual price distribution of each market. We observe that, indeed, the minimum price declines with the number of stations whereas the maximum price increases with N. This is also true of the 10^{th} and 90^{th} quantiles, but perhaps less noticeable in other regions of the support of the residual price distribution. The magnitudes of these changes may not seem large – at most 4 cents and usually fractions of one cent– but this magnitude should be evaluated against the relatively small price dispersion of gasoline prices, which amounts to 1.07 cent for residual prices (see Figure 4).

Of course, these average statistics cannot be used to infer the effect of changes in N on the distribution of prices. The minimum and maximum prices, being order statistics, tend to decrease and increase, respectively, with the sample size. This implies that there is a built-in tendency for the extreme prices to vary systematically with N (which is highly correlated with, but not equal to, the sample size). In addition, there are many factors that affect both N and prices and, if these are not controlled for, part of their effect is attributed to changes in N. Income, for example, affects the number of stations and prices through the relationship between income and willingness to pay

and/or shopping behavior. One would expect the number of stations and prices to be higher in markets with higher incomes inducing a spurious positive relationship between prices and N. In order to control for the effect of sample size and of other confounding factors we proceed to a multivariate regression analysis of the data.

4 Empirical strategy

Let $w_m = (v_m, \mu_m, c_m)$ where $\mu_m = (\mu_{1m}, \mu_{2m}, \ldots, \mu_{S-1m})$ gathers information on consumers' "shopping behavior" in market m. As explained in Section 2, both w_m and N_m determine the equilibrium price distribution. We use residual prices in our analysis so that we now let p_{it} refer to the residual price of station i in day t. In each market m, we have n_m (station-day) observations on residual prices p_{it}. Suppose we observe market characteristics (N_m, w_m). We assume that the sample of residual prices in market m is randomly drawn from the same distribution $F_{N_m}(p) = F(p|N_m, w_m)$. This accords with the way in which prices were collected. Let $q_m(\tau)$ be the τ^{th} quantile of residual price p_{it} in market m. $q_m(\tau)$ is a function of (N_m, w_m), the determinants of the price distribution. By analyzing how N affects $q_m(\tau)$ at different values of τ, we learn about the effect of changing N on the distribution of prices.

In order to do this we first estimate $q_m(\tau)$ by the $[n_m\tau]^{th}$ smallest value among all n_m observations in market m. We denote this estimator by $\widehat{q}_m(\tau)$. This estimator has an asymptotic normal distribution with expected value $q_m(\tau)$ and variance $\frac{\tau(1-\tau)}{n_m f_{N_m}(q_m(\tau))^2}$, where f_{N_m} is the conditional density of p_{it} given (N_m, w_m). In a second step we regress $\widehat{q}_m(\tau)$ on N_m and (proxies for) w_m using the municipality-level data. This provides us with an estimate of the effect of the number of stations on prices. We run a separate regression for each chosen value of τ.[36]

An alternative estimator of the effect of the number of stations on prices could be obtained by estimating the quantile functions directly with the station-level price data p_{it} using a standard "quantile-regression" procedure. One should note, however, that although the two estimators are not numerically identical in finite samples, they

[36] See Chamberlain (1994) and Bassett, Tam and Knight (2002) for examples of this 2-step approach.

are first-order asymptotically equivalent.[37] We do not adopt the quantile regression approach for a number of reasons. First, the regressors (N, w) vary only at the level of the municipality, and this procedure might underestimate standard errors, even if the standard errors were clustered at the municipality level. Because market-level regressions are based on a much smaller number of observations (about 440 observations) than station-level regressions (about 31,000 observations), the former procedure generates more conservative standard errors.[38] Second, the procedure based on station-level data gives more weight to the largest municipalities because the number of sample observations (n_m) increases with the number of stations (N_m). If the effect of the number of stations is weaker in the largest municipalities where $N > S$, procedures based on station-level data might underestimate the effect of competition on prices. We want to give equal weights to all markets so as to be able to interpret the estimated coefficient as the effect of changing N on the price distribution of a market chosen at random.[39] Finally, it is known than the normal approximation is not adequate for extreme quantiles and extreme values (e.g., the minimum and maximum) and this can therefore affect the asymptotic distribution of the estimators based on station-level data. With market-level data, however, we are estimating an average of the extreme values across markets so that a normal approximation applies when the number of markets is large. We therefore use unweighted market-level data to estimate the mean quantiles.

Regarding functional form, we make a separability assumption on the conditional expectation function between N and w and specify the N part in natural logs, i.e.,

$$E\left[\widehat{q}_m(\tau)|N_m, w_m\right] = \beta_0 + \beta_N \ln N_m + h(w_m)$$

[37] If the function being estimated is the mean then a weighted regression of market level data produces the same estimates as those obtained from station level data. But this is not the case for quantiles. In our data, the store-level and weighted market level estimates are quite close to each other (i.e., well within 1 standard deviation of each other), when the weights are the number of observations in each market. This was also observed by Basset, Tam and Knight (2002) in their study of ACT scores by school. See also Knight (2002) for the asymptotic equivalence results and for a comparison of both estimators using simulated data.

[38] An argument made by Guryan and Charles (2008).

[39] The station-level estimates can be interpreted as the effect of changing N on the price distribution faced by a consumer chosen at random. Although this is of interest and certainly important for welfare analysis, it is not the focus of our paper.

The justification for the logarithm specification is that the marginal effect of N on a price quantile is decreasing in N. That is, the effect of N is stronger when N is small than when N is large. This accords with the model outlined in Section 2 (see Figure 2) and with the preliminary empirical evidence presented in Figure 7. The log specification is a parsimonious way of achieving this. In any case, we will check the robustness of our conclusions to alternative functional forms.

The most important econometric problem is that w is unobserved and likely to be correlated with N because w is one of the determinants of N via the zero-profit condition (6). Thus, ignoring the term $h(w)$ and treating it as error will bias our estimates of β_N. We approach this problem in two ways. First, we use an array of covariates x to proxy for w and, secondly, we use instruments for N to deal with the remaining correlation.

Specifically, we take a linear projection of $h(w)$ on x,

$$h(w) = x\pi + r \text{ with } Cov(r, x) = 0$$

which results in

$$E\left[\widehat{q}_m(\tau)|N_m, x_m, r_m\right] = \beta_0 + \beta_N \ln N_m + x_m\pi + r_m \tag{7}$$

The problem with using proxies for $h(w)$ is that there is no guarantee that $\ln(N)$ will be uncorrelated with the unobserved r in equation (7). The correlation between $\ln(N)$ and the residual heterogeneity r, however, need not be strong if x includes the main determinants of w. That is, controlling for sufficient municipality-level characteristics can potentially ameliorate this endogeneity problem. This is the reason why the availability of economic, geographic and demographic data at the municipality level strongly favors identifying markets with municipalities.

Nevertheless, because the proxies are not perfect, the omitted variable bias is never completely eliminated. It is therefore important to generate exogenous variation in N in order to be able to consistently estimate its causal effect on y. Economic theory is useful here because it readily suggests an appropriate instrument for N. As shown in (6), in the long-run market equilibrium with free entry, the number of stations is determined by L, E and w so that, as discussed in Section 2.2, L and E are natural instruments for

N. We need to assume, however, that E and L are exogenous variables in equation (7) in the sense that they are mean-independent of r, conditional on x, i.e.,

$$E(r|E, L, x) = 0 \tag{8}$$

This identifying assumption says that among markets with the same observed characteristics x, variations in population size and entry costs are not associated with w, i.e., with the willingness to pay for gasoline and with shopping behavior. If, for example, more affluent municipalities have higher willingness to pay, higher entry costs and lower population then this assumption would be violated if we do not include measures of income or wealth among the controls x. As usual, the strength of this assumption depends on what is included in x.

For the number of consumers L to be a good instrument, the marginal cost must be independent of market size. In connection with this, we note that variable costs in gasoline retailing are mostly driven by the cost of gasoline. The typical brand in the Netherlands buys its gasoline from the Amsterdam-Rotterdam-Antwerp (ARA) spot market (this is true even for Shell which sells much more gasoline than it produces). The ARA market is a centralized marketplace where price discrimination mechanisms such as quantity discounts are unfeasible due to the anonymity of the traders. Therefore, it is reasonably safe to assume that most gas stations in the Netherlands face similar wholesale gasoline prices irrespective of the population level or density in the areas where the stations are located. As will be seen in the next section, our overidentification tests confirm this assumption.

The standard errors of the estimators need to account for the heteroskedasticity induced by the sampling error in estimating the quantiles, $\frac{\tau(1-\tau)}{n_m f_m(q_m(\tau))^2}$. Instead of estimating the density function we use (White) standard errors that are robust to arbitrary heterogeneity.[40]

[40] As a check we also bootstrapped the standard errors of the 2SLS regressions by resampling from the $(\widehat{q}_m(\tau), N_m, w_m)$ data. These bootstrapped standard errors, based on 1000 replications, are between 8 and 20 percent (15 percent on average) higher than the White standard errors. We chose to use the latter because they are easier to compute and do not alter any of our conclusions regarding statistical significance.

Regarding estimation of the effect of N on the minimum and maximum price we proceed in the same way as we did for the quantiles except that we add n_m to the list of regressors. We do this because the sample minimum and maximum are monotonic functions of the sample size n_m, while n_m is correlated with the number of stations in the market (the simple correlation between n_m and $\ln(N_m)$ is 0.72). In this way we ensure that the estimate of β_N does not reflect the built-in correlation between extreme prices and sample size.[41]

5 Empirical Results

Table 4 presents estimates of several variants of equation (7). The regressions for each price statistic are run separately since there are no efficiency gains to joint estimation when the regressors are the same across equations. Panel A presents OLS estimates of regressing a price statistic on $\ln(N)$ only. These regressions are based on 439 observations (municipalities) because residual prices could not be computed for one municipality (Reiderland) where only one station has prices in only one day. The effect of N is negative for the lower price statistics and positive for the higher ones, as predicted by the theoretical model. The more extreme the price statistics, the more significant are the effects of N. The median price also decreases with N but this effect is marginally significant. As N increases, say from 1 to 2 stations, the minimum price is estimated to decrease by 0.35 cents ($-0.508 \times \log 2$), while the maximum price is estimated to increase by 0.26 cents. These are not small changes relative to the standard deviation in residual prices (1.07 cent).

In Panel B we add 39 provincial dummies to control for unobserved time-invariant effects at the regional level.[42] These regional effects are always jointly significant at the 1 percent level (also in panels C and D). The estimated coefficients of $\ln(N)$ increase

[41] Recall that the sample size n_m depends on the number of stations sampled and on the number of days each station appears in the sample, while N_m is the total number of stations in market m and comes from a different data source.

[42] There are 40 regional areas (known as COROP areas) in the Netherlands. Each regional area comprises several municipalities.

somewhat, particularly for prices in the middle of the distribution.

In Panel C we add proxies for consumers' reservation values and for shopping behavior. We do not directly proxy for c because, as mentioned above, variable production costs are quite similar across markets. The regressions in panel C and D are estimated on 423 markets because of missing data on some of the covariates. The reasons for missing covariate data are unrelated to the price of gasoline and therefore there is no risk of sample selection bias. Indeed, Panel B was reestimated for the sample of 423 observations used in Panel C and D and the estimated coefficients are very similar to those reported in the table.

Perhaps among the main determinants of the willingness to pay and of shopping behavior for gasoline is income. We therefore include average household income among the proxies for both v and μ. Because of income and substitution effects we expect this variable to be positively correlated with the willingness to pay, and negatively correlated with the proportion of fully-informed consumers. Thus, income should positively affect prices. We also include the share of cars registered to business (out of total cars in the municipality), which should be positively correlated with the willingness to pay for gas and therefore also affect prices positively.

An additional set of controls is related to the geographic characteristics of markets. The distribution of consumer types may vary with the geography of the market. Consumers' shopping behavior may be different in a geographically small, interconnected municipality than in a large, spatially-spread municipality. We therefore add controls for the total area of the municipality (in km^2), the area that is land (also in km^2), the share of land that is built (urbanized) and the share that is agrarian (the remainder is land for recreation and forests), and the kilometers of roads within the municipality borders.[43]

We also add the sample size n_m in each market to all the regressions in panel C, not only to the minimum and maximum price regressions. In this way we control for possible sample size effects on the estimation of the quantiles. In practice, however, the coefficient of n_m is not significantly different from zero, and excluding the sample size

[43] We would also like to have a measure of the distance between stations in a municipality but unfortunately we do not have these data.

from the quantile regressions does not affect the estimates of β_N.

The effect of adding these additional regressors is to lower the estimates of β_N, particularly for the lower prices. In addition, the precision of the estimates decreases because of the additional estimated parameters, rendering most of the estimated β_N's, except in the minimum and maximum price regressions, not significantly different from zero.

Our final set of regressions in panel D uses 2SLS to eliminate potential biases from unobserved common determinants of prices and of N. An additional reason for using 2SLS is that the number of stations is likely to be measured with error because our data for N correspond to stations operating during August 2007, while our price data were collected in May 2006.

Free entry and a zero profit condition predict a positive relationship between the number of stations in the market and population size L, and a negative relationship between N and entry costs. We do not have data on market-specific entry costs but we have data on the level of municipality taxes imposed on business real estate and use these as a measure of E.[44] First-stage regression results appear in Table 5. In columns (1) we regress the number of stations (in logs) on population and the tax rate (both in logs), while in column (2) we add 39 provincial dummies.[45] Both instruments have coefficients with the predicted sign and are significantly correlated with $\ln(N)$. As controls are added to the regression, in columns (3) and (4), the effect of taxes is halved and loses it statistical significance. Nevertheless, in all regressions, the F-test for joint significance of population and taxes is very high indicating that these are strong instruments. Column (4) corresponds to the first-stage in the 2SLS procedure used in Table 4.[46]

Panel D in Table 4 reports the 2SLS estimates of β_N using the same specification

[44] Tax rates vary between 1.5 and 18 percent across municipalities with an average of 7.1 percent.

[45] Using the tax rate in levels instead of logs works equally well. We treat L and E symmetrically as it would be suggested by a logarithmic approximation to the zero profit condition (6).

[46] There are 424 municipalities with data on all covariates but one of them (Reiderland) does not have residual price data. Thus, the total number of observations used in the price regressions in Table 4 is 423.

as in panel C. The 2SLS results are in line with the previous estimates but they are larger in absolute value than the OLS estimates in panel C; these differences are quantitatively important. On the one hand, because unobserved determinants of prices are likely to be positively correlated with the number of stations, using 2SLS should increase the absolute value of the negative estimates of β in the minimum and low price regressions and decrease the positive estimates of β in the maximum and high price regressions. On the other hand, in the presence of measurement errors in N, OLS estimators are biased towards zero in all the regressions. Since 2SLS also removes the correlation with the measurement error, 2SLS estimates of the coefficient in the lower price regressions should be, on both accounts, more negative than the OLS estimates. This is indeed what we observe in panel D. For the higher price regressions, the biases in OLS due to omitted variables and to measurement error work in opposite directions and it is therefore not possible to predict in which direction the estimates should change with 2SLS.

The marginal effects of an increase in the number of stations from N to $N+1$ is

$$\beta_N \ln\left(\frac{N+1}{N}\right)$$

These marginal effects are plotted in Figure 8 for all the 7 price statistics for $N = 1, \ldots, 50$, along with a 2 standard deviation band. We observe that for small values of N, the marginal effects are indeed positive for the higher prices and negative for the smaller prices. These effects are also significantly different from zero. The logarithmic specification implies that marginal effects converge to zero as N increases.

Using these estimates we find, for example, that adding 4 additional gas stations to a single-station market would lower the minimum price of a liter of Euro 95 by 0.93 cents ($-0.58 \times \log 5$) and increase the maximum price by 0.83 cents. Recall that the standard deviation in the residual price distribution is 1.07 cents and therefore these estimated effects are quantitatively significant relative to the dispersion in prices.

The effect of the covariates would be identically zero if we were estimating the mean residual price. Since we are estimating the extremes and quantiles of the price distribution these effects need not be zero. Nevertheless, the estimated coefficients of the control variables (not reported) are usually not significantly different for zero (in-

dividually and jointly). This is driven in part by the inclusion of regional dummies in the regression which correlate with the municipality-level characteristics. Indeed, if the provincial dummies are removed from the regressions in panel D, the controls are jointly significant in four of the seven regressions (while the estimates of β_N are virtually unaffected). Moreover, as we will see later, when the actual instead of the residual prices are used to compute the extreme prices and the quantiles, the controls are significant in all regressions (see bottom panel in Table 6).

In sum, the empirical findings suggest that as the number of stations in the market increases, the low prices tend to decrease while the high prices tend to increase. This characterization of the effect of competition on prices accords with most of the theoretical predictions of the model presented in Section 2.[47] Although we cannot check what happens to the mean price as N increases, we found that the median price is lower in markets with more stations. If the mean and median do not differ much, then this is a result that is at odds with the theoretical prediction of Claim 1. This asymmetric effect of a change in the number of firms may have been remarked theoretically but, to the best of our knowledge, has not been analyzed empirically. The welfare implications of such asymmetry will be studied in Section 6. We first perform a set of robustness checks.

5.1 Robustness checks

Underlying the results in Table 4 are a series of specification assumptions which, if incorrect, could lead to biased estimates of β_N. We now examine these assumptions in greater detail and verify that our 2SLS results in Table 4 are robust to departures from our baseline specification.

First, we note that the absolute value of the estimated coefficients β_N in both extreme price regressions – -0.58 and 0.52 – are an order of magnitude larger than the corresponding estimates in the quantile regressions. We are concerned that this may reflect our poor attempt to control of the effect of sample size on extreme prices. Although this effect is highly non-linear, we just used n_m linearly in the regressions.

[47] And, more generally, with models where some consumers have imperfect information about prices and observe (perhaps through search) different number of prices.

However, adding a quadratic of the sample size, n_m^2, does not affect the estimated results – in fact, it makes them slightly stronger (results not reported).[48] We can gain further understanding of how the improper control of sample size may bias our estimates by estimating β_N for quantiles closer to the extreme prices, namely, the 1, 2,98 and 99 quantiles. These quantiles are not that sensitive to sample size as the minimum and maximum prices are. Thus, assuming that the $\beta'_N s$ are the same in the extreme prices and nearby quantiles regressions, the difference in their estimated coefficients is indicative of the effect of sample size on the estimates. The estimated $\beta'_N s$ for the 1^{st} and 2^{nd} quantiles are, respectively, -0.59 and -0.32, while for the 99^{th} and 98^{th} quantiles they are, respectively, 0.45 and 0.32. In all cases these estimates are highly significant. Thus, although we cannot rule out some bias due to sample size, fully accounting for this bias is not likely to alter our qualitative conclusions.

We next address functional form issues. Although entering the number of stations in log form is parsimonious as well as theoretically appealing – the model implies that there is no effect of N on prices when $N > S$ – it may be practically restrictive. We allowed the coefficient of $\log N$ to change for $N \geq N_0$, for various levels of N_0 ($N_0 = 2, \ldots, 11$), but the interaction term was never significantly different from zero. We also added the square of $\ln(N)$ to the regression to allow for more flexibility in the marginal effect but this term was never significantly different from zero (its p-value ranged between 0.34 and 0.99) except in the maximum price regression (p-value 0.06).[49] Adding $(\ln(N))^2$ made the coefficient of $\ln(N)$ also insignificant. This is not surprising because $\ln(N)$ and $(\ln(N))^2$ are highly correlated; their simple correlation coefficient is 0.95. Although the individual parameters cannot be precisely estimated, the marginal effects track very closely the marginal effects estimated from the regression in panel D of Table 4 except, perhaps, for those of the 75^{th} quantile price (see Figure 9).

We could avoid making strong functional form assumptions if we allow for the effect of N to vary non-parametrically with N. This can be achieved by using dummy

[48] We did not enter n_m in logs because of the even higher collinearity with $\ln(N)$ – simple correlation coefficient of 0.89.

[49] We added the squares of log population and log tax to the list of instruments.

variables for each value of N. The problem here is that N takes on 36 distinct values and the corresponding dummies would still be endogenous. Even if we had the large number of instruments required (or use a control function approach), this approach is not practical with the sample size we have. We therefore group the number of stations into 4 size groups and add dummies corresponding to these groups. The four groups are defined as markets with 1 and 2 stations – the baseline group –, markets with 3-6 stations, markets with 7-10 stations and markets with more that 11 stations. In order to address the endogeneity of these group dummies we follow the procedure suggested by Wooldridge (2002, p. 623) and first estimate a probit equation for the probability that the number of stations in a market is in a given size group. We run a separate regression for each size group and compute the predicted probability of belonging to a size group. In this regression we include the same regressors as in the first stage of the 2SLS estimator in panel D of Table 4. We then run the regressions as in panel D using the predicted probabilities as instruments for the endogenous group dummies.[50]

Results of this two-stage 2SLS estimation are presented in the top panel of Table 6, where the coefficients represent the change in price in a given group size relative to the preceding group size. We see that markets with 3-6 stations have lower low prices but higher high prices than markets with 1-2 stations. Markets with 7-10 stations exhibit the same pattern, relative to markets with 3-6 stations, but the effects are of lower magnitude and less significant. Finally, the estimates indicate that prices in markets with 11 or more stations are not significantly different from prices in markets with 7-10 stations. These findings accord, at least in a qualitative sense, with the marginal effects depicted in Figure 8. In sum, altering the simple functional form used in Table 4 would not change our conclusions regarding the asymmetric effect of competition.

Next, we examine what happens to our estimates of β_N when the dependent variables are based on the actual (raw) prices and not on the prices net of station and day effects. We do not pool observations over time because the wholesale price may be chang-

[50] Wooldridge (2002, p. 623) shows that the first-stage probit regressions need not be correctly specified, and that inference based on the standard errors of the 2SLS procedure is correct even though the instruments are generated in a previous step.

ing over the sample period and therefore compute the minimum, mean, maximum and other quantiles of the price distribution for each market and for each day. We cannot control for unobserved station-level effects because doing so wipes out all market-level regressors, including $\ln(N)$. We now have 7091 market-day observations. We estimate the same model as in panel D of Table 4.[51] The results using raw prices, which now include a regression for the mean price, are in the bottom panel of Table 6. The mean and median price do not appear to be significantly affected by the number of stations in the market but the low and high prices are. The estimated parameters follow the same pattern as in Table 4 but are much stronger than the estimates based on residual prices. This is to be expected simply because the extreme prices and quantiles based on raw prices are larger than those based on residual prices. Because the stronger coefficients in the raw price regressions reflect a scaling effect it is important to base our analysis on prices net of station-specific components.

In contrast to the regressions in Table 4 where the dependent variables were based on residual prices, the covariates are now jointly significantly different from zero in all regressions. The sample size variable is negative and significant in the minimum price and the 10^{th}quantile regressions while it is positive and significant in the 90^{th} quantile and the maximum price regressions; it is not significantly different from zero in the other regressions. Average household income and the share of cars registered to business always have positive coefficients but only the latter are significant. The geographic controls are significant in four of the seven regressions.

The number of stations in the market is defined as the number of stations in the municipality. It may well be that the "relevant" number of stations affecting prices in a market includes the stations in neighboring municipalities. In order to examine this possibility we computed, for each market m, the number of stations in all the municipalities sharing a border with market m and added the log of this variable to the

[51] The only differences with the specification in Table 4 are that the dependent variable and the sample size regressor change over days, and that we added day dummies to control for the effect of the day in the month. The other regressors are constant over time. Standard errors were clustered at the market level to allow for arbitrary serial correlation and heteroskedasticity.

basic model. The results appear in the top panel of Table 7. We drop 3 municipalities that are islands and therefore have no neighbors. Essentially, the number of gas stations in neighboring markets has a much smaller effect on prices that the own number of neighbors and, in all cases, this effect is not significantly different from zero. Importantly, the estimated β_N's are almost unaffected by the inclusion of the number of neighboring stations in the regression.[52]

We also re-run the regression excluding the four largest markets in the Netherlands (Amsterdam, Rotterdam, 's-Gravenhage, and Utrecht). Since these cities represent only 4 observations we do not expect to obtain very different results. And, indeed, the estimated coefficients based on restricted sample of smaller cities are very similar to those in panel D of Table 4.[53] In this vein, recall that the theoretical model presented in Section 2 does not examine markets with a single station. There are 16 municipalities where $N = 1$. Removing these observations from the regressions makes each of the estimated coefficients even stronger.[54]

Finally, our findings are not restricted to a particular gas product (Euro 95). The other popular product in gas stations is, of course, Diesel. The bottom panel in Table 7 replicates the regression in panel D of Table 4 for residual Diesel prices. The estimated coefficients are remarkably similar to those from the Euro 95 regressions.

6 Welfare implications

The evidence presented above points to significant differences in the way increased competition – increased number of gas stations – affects different parts of the price distribution. Whether consumers are successful in paying the lower prices depends on their shopping behavior. Increased competition is likely to favor those consumers observing many prices

[52]Because provincial dummies pick up regional effects, we treat the number of neighboring stations as exogenous in the price regressions. The overidentification tests support this assumption.

[53]These new estimates, in the order in which they appear in Table 4, are: -0.669 (0.231), -0.131 (0.0595), -0.0435 (0.0395), -0.0568 (0.0337), 0.0276 (0.0569), 0.105 (0.0721) 0.523 (0.145).

[54]The estimated coefficients, in the order appearing in Table 4, are -0.712, -0.099, -0.0766, -0.0641, 0.0900, 0.1472, 0.5473 and are slightly more significant as those in panel D of Table 4.

but may hurt those observing only a few prices. It is therefore not obvious – in contrast to the full-information model – that all consumers benefit from more competition. In this section we study this issue in detail and quantify the welfare gains from increased competition for different group of consumers.

As explained in Section 2, consumers observing $1 \leq s \leq N$ prices, pay $y_s \equiv Min\{p_1, p_2, ..., p_s\}$ each time they purchase, where $p_1, p_2, ..., p_s$ are i.i.d. random draws from F_N. On average, these consumers pay

$$E[y_s] = E[Min\{p_1, p_2, ..., p_s\}]$$

where the expectation is taken using F_N, the equilibrium price distribution in a market with N stations.[55]

We estimate $E[y_s]$ for each market as follows. We draw s residual prices with replacement from the sample of n_m residual prices observed in each municipality (pooled over gas stations and days).[56] We take the minimum of the s prices and store it. We repeat this 10,000 times and compute the average of the 10,000 stored minimum prices. This average is our estimate of $E[y_s]$ for each $s = 1, 2, \ldots, N$ in each market (characterized by a given N). That is, we obtain N estimates of $E(y_s)$ in each market corresponding to the expected price paid by different consumers observing, respectively, $s = 1, \ldots, N$ prices.

There are two dimensions of these estimates that are of interest. First, as consumers observe more prices in a given market, the price they end up paying should be lower on average. That is, broader price exposure should result in lower prices paid. We clearly see this in Figure 10 where we plot the estimates of $E[y_s]$ in each market against s, as well as the predicted value of a locally weighted regression of the estimate of $E[y_s]$ on s. The gains from being better informed – the difference in expected price paid as s increases by $1-$ are positive in 99.4 percent of the observations.

[55] That is, $E[y_s] = \int ps(1 - F_N(p))^{s-1} f_N(p)dp$.

[56] We report estimates based on prices drawn with replacement because sample sizes are relatively small. Since, in reality, consumers do not sample with replacement we also replicated our calculations when prices are drawn without replacement; the results are practically identical.

Although the path of expected prices paid in each market declines with s, Figure 10 points out that there is substantial heterogeneity in the price paid for given s across markets. This heterogeneity is a reflection of the different price distributions across markets having different number of stations. This is precisely the other issue of interest – and the focus of this Section – namely, the relationship between N and the price paid, $E[y_s]$, for given s.

To address this issue, we regress our estimate of $E[y_s]$ on $\ln(N)$ and on the other controls used in the previous regressions. We run a separate 2SLS regression for various values of s. Note that when $s = 1$ the estimate of $E(y_1)$ is the mean price in the market which is zero by construction. We therefore present, in Table 8, the estimated effect of $\ln(N)$ on $E(y_s)$ for $s = 2, \ldots, 7$, i.e., for consumers that observe up to 7 prices. Note also that as s increases, the number of observations declines because there are fewer municipalities with N above s (see Table 1) and the parameters are not precisely estimated for $s \geq 5$.

Two results in Table 8 are noteworthy. First, the estimates are all negative. A negative coefficient means that the prices paid by consumers decrease as the number of competitors increases. Although, as shown in Section 2, it is theoretically plausible that some consumers may be worse off when competition increases, in this application all types of consumers benefit from increasing the number of gas stations. Second, this negative effect of N increases with s up to, and including, $s = 4$ but stabilizes thereafter. This means that the gains from increased competition – in terms of price reduction – are maximal for consumers observing 4 prices. Entry of additional stations does not result in gains for consumers who observe 5 or more prices. In other words, the magnitude of the welfare improvement depends on shopping behavior and is larger for those consumers that observe more prices.

In Figure 11, we use the estimates in Table 8 to plot the marginal effect of N on the expected price paid for $s = 2, \ldots, 7$ and for $N \leq 15$ (marginal effects are not different from zero for larger values of N). We observe that as competition increases the price paid by all types of consumers declines in markets with up to 6-8 gas stations; thereafter the expected price paid remains constant. The decline, however, is about twice as large

for consumers that observe 4-5 prices than for consumers that observe only 2 prices.

Finally, we note that the welfare analysis was based on residual prices. It is conceivable that the stations' characteristics and/or their productivity also respond to competition and that this response is reflected in the stations' mean prices. By focusing on the residual prices we may be missing some important effects of competition. This does not happen to be the case in our data since the results in the bottom panel of Table 6 point to non-significant effects of the number of stations on the mean price of gasoline.

7 Conclusions

In markets where the amount of price information varies across consumers, prices are typically dispersed in equilibrium. An increase in the number of firms usually affects each seller's pricing strategy and this in turn alters the entire distribution of equilibrium prices. Traditionally, empirical research has focused on estimating the impact of competition on the mean and variance of prices.[57] Although this is certainly useful, these statistics are not sufficient to perform a detailed welfare analysis because competition can affect different parts of the price distribution in opposite directions.

This paper has tried to fill this gap. We examined how the distribution of gasoline prices (the minimum, median and maximum prices, as well as other quantiles) in the Netherlands changes with the number of competitors in the market. We used population size and local taxes as instruments for the number of gas stations. We found that as competition – the number of gas stations – increases the distribution of prices spreads out, with the low prices going down and the high prices going up. Consequently, competition has an asymmetric effect on prices.

This result has important welfare implications because when some prices increase and others decline, the price actually paid by consumers will depend on their shopping behavior. All (hypothetical) consumers in our data, irrespective of whether they are informed about one or more prices, benefit from an increase in the number of stations.

[57] In our data, the standard deviation of residual prices increases with the number of stores. The 2SLS coefficient of $\ln(N)$ in a regression specification similar to those in panel D of Table 4 is 0.086 (s.e. 0.033).

The magnitude of the welfare gain, however, is greater for those consumers that observe more prices. As a result, an increase in competition has a positive but unequal effect on the welfare of consumers.

Since price dispersion is prevalent in many markets, we believe the paper has a general message that goes beyond the present application to the gasoline market in the Netherlands. The price effects of competition-enhancing policies (e.g., industry deregulation, trade liberalization, etc.) are not as straightforward as one may be led to believe based on standard oligopoly theory. As a result, welfare implications are not obvious either. In fact, we have shown, theoretically and empirically, that increased competition can have unequal effects among consumers; some consumers may even experience declines in their welfare as a result of higher prices.

A complete welfare analysis, however, would require a mapping between shopping behavior and socio-economic characteristics of interest and analyzing other dimensions of consumer welfare. For example, if consumers that observe only a few prices are high-income consumers (whose value of time is higher) then these consumers benefit less from competition than low-income consumers. These benefits refer only to price changes and do not take into account other welfare effects of competition such as increases in the variety of goods offered to consumers.

Lastly, although our empirical work is motivated by a particular theoretical framework, we think the empirical findings reported in the paper are of interest on their own right and, if verified in other data sets, they should be taken into account when formulating theoretical models of pricing in oligopolistic markets.

References

[1] Armstrong, Mark (2008), "Interactions between Competition and Consumer Policy," *Competition Policy International* 4(1), 97-147.

[2] Baye, Michael. R., John Morgan and Patrick Scholten (2004), "Price Dispersion in the Small and in the Large: Evidence from an Internet Price Comparison Site", *The Journal of Industrial Economics,* 52(4), 463-496.

[3] Baye, Michael. R., John Morgan and Patrick Scholten (2006), "Information, Search, and Price Dispersion, chapter 6 in *Handbook on Economics and Information Systems Volume 1,* (T. Hendershott, Ed.), Amsterdam: Elsevier.

[4] Bassett,Gilbert, Mo-Yin S. Tam and Keith Knight (2002), "Quantile Models and Estimators for Data Analysis", *Metrika*, 55, 17-26.

[5] Barron, John M., Taylor, Beck A. and John R. Umbeck (2004), "Number of sellers, average prices, and price dispersion", *International Journal of Industrial Organization* 22(8-9), 1041-1066.

[6] Borenstein, Severin and Nancy Rose (1994), "Competition and Price Dispersion in the U.S. Airline Industry," *Journal of Political Economy*, 102, 653-683.

[7] Chamberlain, Gary (1994), "Quantile regression, censoring and the structure of wages", chapter 5 in *Advances in Econometrics, Sixth World Congress*, edited by Christopher Sims, Cambridge University Press.

[8] Chandra, Ambarish and Mariano Tappata (2008), "Price Dispersion and Consumer Search in the Retail Gasoline Market", http://strategy.sauder.ubc.ca/tappata/research.htm.

[9] Janssen, Maarten and José Luis Moraga-González (2004), "Strategic Pricing, Consumer Search and the Number of Firms", *Review of Economic Studies*, 71(4), 1089-1118.

[10] Guryan, Jonathan and Kerwin Charles (2008), "Prejudice and Wages: An Empirical Assessment of Becker's The Economics of Discrimination", *Journal of Political Economy*, October 2008, 116(5), 773-809.

[11] Haile, Philip A., Han Hong, and Mathew Shum (2003), "Nonparametric Tests for Common Values in First-Price Sealed-Bid Auctions", http://www.econ.yale.edu/~pah29/npmcv.pdf.

[12] Hosken, Daniel S., Robert S. McMillan and Christopher T. Taylor (2008), "Retail Gasoline Pricing: What Do We Know?", *International Journal of Industrial Organization*, 26(6), 1425-1436.

[13] Hu, Yingyao and Matthew Shum (2008), "Estimating First-Price Auction Models with Unknown Number of Bidders: a Misclassification Approach", http://www.econ.jhu.edu/people/shum/papers/montecarlo.pdf.

[14] Knight, Keith (2002), "Comparing conditional quantile estimators: first and second order considerations",http://www.utstat.utoronto.ca/keith/home.html.

[15] Lach, Saul (2002), " Existence and Persistence of Price Dispersion: an Empirical Analysis", *Review of Economics and Statistics*, August, 433-444.

[16] Lewis, Matthew (2008), "Price Dispersion and Competition with Differentiated Sellers", *Journal of Industrial Economics* 56 (3), 654-678.

[17] Stahl, Dale O. (1989), "Oligopolistic Pricing with Sequential Consumer Search", *American-Economic-Review*, 79(4), 700-712.

[18] Varian, Hal (1980), "A Model of Sales", *American Economic Review*, 70, 651-659.

[19] Waterson, Michael (2003), "The Role of Consumers in Competition and Competition Policy," *International Journal of Industrial Organization* 21, 129-50, 2003.

[20] Wildenbeest, Matthijs R. (2008), An empirical model of search with vertically differentiated products, mimeo.

[21] Wooldridge, Jeffrey M. (2002), *Econometric Analysis of Cross Section and Panel Data*, The MIT Press, Cambridge.

Appendix

Proof of Proposition 1. (A) *Let p_N be a random variable with distribution $F_N(p)$; likewise, let p_{N+1} be a random draw from $F_{N+1}(p)$. Then $E[p_{N+1}] > E[p_N]$.*

We first note that

$$E[p_N] = \int_{\underline{p}}^{v} p dF_N(p) = v - \int_{\underline{p}}^{v} F_N(p)dp = \int_0^1 p(F_N)dF_N$$

Using equation (2) we have that

$$E[p_N] = c + \int_0^1 \frac{(v-c)\mu_1}{\sum_{s=1}^{S-1} s\mu_s x^{s-1} + N\gamma x^{N-1}} dx$$

We are interested in the sign of $E[p_{N+1}] - E[p_N]$. We can write:

$$\begin{aligned}\frac{E[p_{N+1}] - E[p_N]}{(v-c)\mu_1} &= \int_0^1 \left(\frac{1}{\sum_{s=1}^{S-1} s\mu_s x^{s-1} + (N+1)\gamma x^N} - \frac{1}{\sum_{s=1}^{S-1} s\mu_s x^{s-1} + N\gamma x^{N-1}} \right) dx \\ &= \int_0^1 \frac{\gamma x^{N-1}(N-(N+1)x)}{\left(\sum_{s=1}^{S-1} s\mu_s x^{s-1} + (N+1)\gamma x^N\right)\left(\sum_{s=1}^{S-1} s\mu_s x^{s-1} + N\gamma x^{N-1}\right)} dx\end{aligned}$$

This last integral takes positive values only when $x < N/(N+1)$. Therefore we have

$$\begin{aligned}\frac{E[p_{N+1}] - E[p_N]}{(v-c)\mu_1} &= \int_0^{\frac{N}{N+1}} \frac{\gamma x^{N-1}(N-(N+1)x)}{\left(\sum_{s=1}^{S-1} s\mu_s x^{s-1} + (N+1)\gamma x^N\right)\left(\sum_{s=1}^{S-1} s\mu_s x^{s-1} + N\gamma x^{N-1}\right)} dx \\ &- \int_{\frac{N}{N+1}}^1 \frac{\gamma x^{N-1}((N+1)x-N)}{\left(\sum_{s=1}^{S-1} s\mu_s x^{s-1} + (N+1)\gamma x^N\right)\left(\sum_{s=1}^{S-1} s\mu_s x^{s-1} + N\gamma x^{N-1}\right)} dx\end{aligned}$$

Since the denominators of these last two integrals are monotonically increasing in x, we can write:

$$\begin{aligned}&\frac{E[p_{N+1}] - E[p_N]}{(v-c)\mu_1} \\ &> \frac{\int_0^{\frac{N}{N+1}} \gamma x^{N-1}(N-(N+1)x)dx}{\left(\sum_{s=1}^{S-1} s\mu_s \left(\frac{N}{N+1}\right)^{s-1} + (N+1)\gamma\left(\frac{N}{N+1}\right)^N\right)\left(\sum_{s=1}^{S-1} s\mu_s^{s-1}\left(\frac{N}{N+1}\right) + N\gamma\left(\frac{N}{N+1}\right)^{N-1}\right)} \\ &- \frac{\int_{\frac{N}{N+1}}^1 \gamma x^{N-1}((N+1)x-N)}{\left(\sum_{s=1}^{S-1} s\mu_s^{s-1}\left(\frac{N}{N+1}\right) + (N+1)\gamma\left(\frac{N}{N+1}\right)^N\right)\left(\sum_{s=1}^{S-1} s\mu_s^{s-1}\left(\frac{N}{N+1}\right) + N\gamma\left(\frac{N}{N+1}\right)^{N-1}\right)} \\ &= \frac{\int_0^1 \gamma x^{N-1}(N-(N+1)x)}{\left(\sum_{s=1}^{S-1} s\mu_s x^{s-1} + (N+1)\gamma x^N\right)\left(\sum_{s=1}^{S-1} s\mu_s x^{s-1} + N\gamma x^{N-1}\right)} = 0\end{aligned}$$

The proof of part (A) is now complete.

(B) *Let* $y_N = \min\{p_1, p_2, ..., p_N\}$ *where where* $p_1, p_2, ..., p_N$ *are i.i.d. according to* $F_N(p)$. *Then, there exists a number* $\overline{\gamma}(N)$ *such that* $E[y_N]$ *is decreasing in* N *for all* $\gamma < \overline{\gamma}(N)$.

Let $G(y_N) = 1 - (1 - F(y_N))^N$ denote the distribution of y_N. We are interested in

$$E[y_N] = \int_{\underline{p}}^{v} y_N dG(y_N) = v - \int_{\underline{p}}^{v} G(y_N) dy_N = \int_0^1 y_N(G) dG$$

Since $(1 - G(y_N))^{1/N} = 1 - F(y_N)$, then using equation (2) we have that

$$(y_N - c)\left[\sum_{s=1}^{S-1} s\mu_s (1-G)^{\frac{s-1}{N}} + \gamma N (1-G)^{\frac{N-1}{N}}\right] = \mu_1(v-c).$$

Therefore

$$E[y_N] = \int_0^1 y_N(G) dG = c + \int_0^1 \frac{\mu_1(v-c)}{\sum_{s=1}^{S-1} s\mu_s x^{\frac{s-1}{N}} + \gamma N x^{\frac{N-1}{N}}} dx$$

We want to prove that $E[y_N] - E[y_{N+1}] > 0$ for small γ. We can then write:

$$\begin{aligned}
\frac{E[y_N] - E[y_{N+1}]}{\mu_1(v-c)} &= \int_0^1 \left(\frac{1}{\sum_{s=1}^{S-1} s\mu_s x^{\frac{s-1}{N}} + \gamma N x^{\frac{N-1}{N}}} - \frac{1}{\sum_{s=1}^{S-1} s\mu_s x^{\frac{s-1}{N+1}} + \gamma (N+1) x^{\frac{N}{N+1}}} \right) dx \\
&= \int_0^1 \frac{\sum_{s=1}^{S-1} s\mu_s x^{\frac{s-1}{N+1}} - \sum_{s=1}^{S-1} s\mu_s x^{\frac{s-1}{N}} + \gamma(N+1) x^{\frac{N}{N+1}} - \gamma N x^{\frac{N-1}{N}}}{\left(\sum_{s=1}^{S-1} s\mu_s x^{\frac{s-1}{N}} + \gamma N x^{\frac{N-1}{N}}\right)\left(\sum_{s=1}^{S-1} s\mu_s x^{\frac{s-1}{N+1}} + \gamma(N+1) x^{\frac{N}{N+1}}\right)} dx \\
&= \int_0^1 \frac{\sum_{s=1}^{S-1} s\mu_s x^{\frac{s-1}{N}} \left(x^{-\frac{s-1}{N(N+1)}} - 1\right) + \gamma x^{\frac{N-1}{N}} ((N+1) x^{\frac{1}{N(N+1)}} - N)}{\left(\sum_{s=1}^{S-1} s\mu_s x^{\frac{s-1}{N}} + \gamma N x^{\frac{N-1}{N}}\right)\left(\sum_{s=1}^{S-1} s\mu_s x^{\frac{s-1}{N+1}} + \gamma(N+1) x^{\frac{N}{N+1}}\right)} dx
\end{aligned}$$

For convenience, we now separate this last integral as follows:

$$\begin{aligned}
\frac{E[y_N] - E[y_{N+1}]}{\mu_1(v-c)} = &\int_0^1 \frac{\sum_{s=1}^{S-1} s\mu_s x^{\frac{s-1}{N}} \left(x^{-\frac{s-1}{N(N+1)}} - 1\right) dx}{\left(\sum_{s=1}^{S-1} s\mu_s x^{\frac{s-1}{N}} + \gamma N x^{\frac{N-1}{N}}\right)\left(\sum_{s=1}^{S-1} s\mu_s x^{\frac{s-1}{N+1}} + \gamma(N+1) x^{\frac{N}{N+1}}\right)} \\
&+ \int_0^1 \frac{\gamma x^{\frac{N-1}{N}} ((N+1) x^{\frac{1}{N(N+1)}} - N) dx}{\left(\sum_{s=1}^{S-1} s\mu_s x^{\frac{s-1}{N}} + \gamma N x^{\frac{N-1}{N}}\right)\left(\sum_{s=1}^{S-1} s\mu_s x^{\frac{s-1}{N+1}} + \gamma(N+1) x^{\frac{N}{N+1}}\right)}
\end{aligned}$$

We note that the first integral in this expression is always positive, given that x takes values on $[0, 1]$. Let us then analyze the sign of the last integral; for convenience,

let us refer to it as integral I_1. Note that I_1 takes positive values only for values of x above $(N/(N+1))^{N(N+1)}$. Therefore we write:

$$I_1 = -\int_0^{(\frac{N}{N+1})^{N(N+1)}} \frac{\gamma x^{\frac{N-1}{N}}(N-(N+1)x^{\frac{1}{N(N+1)}})dx}{\left(\sum_{s=1}^{S-1} s\mu_s x^{\frac{s-1}{N}} + \gamma N x^{\frac{N-1}{N}}\right)\left(\sum_{s=1}^{S-1} s\mu_s x^{\frac{s-1}{N+1}} + \gamma(N+1)x^{\frac{N}{N+1}}\right)}$$
$$+\int_{(\frac{N}{N+1})^{N(N+1)}}^{1} \frac{\gamma x^{\frac{N-1}{N}}((N+1)x^{\frac{1}{N(N+1)}}-N)}{\left(\sum_{s=1}^{S-1} s\mu_s x^{\frac{s-1}{N}} + \gamma N x^{\frac{N-1}{N}}\right)\left(\sum_{s=1}^{S-1} s\mu_s x^{\frac{s-1}{N+1}} + \gamma(N+1)x^{\frac{N}{N+1}}\right)}dx$$

Notice that the denominator of these integrals are strictly increasing in x. Therefore, setting $x = 0$ in the denominator of the first integral and $x = 1$ in the denominator of the second integral, we have:

$$I_1 > -\int_0^{(\frac{N}{N+1})^{N(N+1)}} \frac{\gamma x^{\frac{N-1}{N}}(N-(N+1)x^{\frac{1}{N(N+1)}})dx}{\mu_1^2}$$
$$+\int_{(\frac{N}{N+1})^{N(N+1)}}^{1} \frac{\gamma x^{\frac{N-1}{N}}((N+1)x^{\frac{1}{N(N+1)}}-N)dx}{\left(\sum_{s=1}^{S-1} s\mu_s + \gamma N\right)\left(\sum_{s=1}^{S-1} s\mu_s + \gamma(N+1)\right)}$$

Integrating we obtain that

$$I > \Omega(N, \mu_1, \mu_2, ..., \mu_{N-1}, \gamma)$$

where

$$\Omega(\cdot) = -\frac{\gamma}{\mu_1^2}\frac{(N+1)(\frac{N}{N+1})^{N(2N+1)}}{(4N^2-1)} + \frac{(N+1)(\frac{N}{N+1})^{N(2N+1)} + 2N^2 - 1}{\left(\sum_{s=1}^{S-1} s\mu_s + \gamma N\right)\left(\sum_{s=1}^{S-1} s\mu_s + \gamma(N+1)\right)(4N^2-1)}$$

Note that $\Omega(\cdot)$ is clearly positive for γ sufficiently small; in fact for $\gamma \to 0$ we just have

$$\lim_{\gamma \to 0} \Omega(\cdot) = \frac{(N+1)(\frac{N}{N+1})^{N(2N+1)} + 2N^2 - 1}{\left(\sum_{s=1}^{S-1} s\mu_s\right)^2 (4N^2-1)} > 0.$$

This shows that $E[min\{p_1, p_2, ..., p_N\}]$ is decreasing in N for γ close to zero. The critical $\overline{\gamma}(N)$ is given by the solution to $\Omega(\cdot) = 0$. Given this solution, $E[min\{p_1, p_2, ..., p_N\}]$ decreases in N for all $\gamma < \overline{\gamma}(N)$. The proof is now complete. We finally note that this is a sufficient condition but not necessary. In fact, numerical simulations of the model show that the results holds for arbitrary γ. The proof of part (B) is now complete.

(C) *Let* $z_N = \max\{p_1, p_2, ..., p_N\}$ *where where* $p_1, p_2, ..., p_N$ *are i.i.d. according to* $F_N(p)$. *Then, there exists a number* $\widehat{\gamma}(N)$ *such that* $E[z_N]$ *is increasing in* N *for all* $\gamma > \widehat{\gamma}(N)$.

Let $H(z_N) = F(z_N)^N$ denote the distribution of z_N. We are interested in

$$E[z_N] = \int_{\underline{p}}^{v} z_N dH(z_N) = v - \int_{\underline{p}}^{v} H(z_N)dz = \int_0^1 z_N(H)dH$$

Since $F(z_N) = H(z_N)^{1/N}$, using the equilibrium equation we have

$$(z_N - c)\left[\sum_{s=1}^{S-1} s\mu_s(1 - H^{\frac{1}{N}})^{s-1} + \gamma N(1 - H^{\frac{1}{N}})^{N-1}\right] = \mu_1(v - c).$$

Therefore we have

$$E[z_N] = c + \int_0^1 \frac{\mu_1(v-c)}{\sum_{s=1}^{S-1} s\mu_s(1 - H^{\frac{1}{N}})^{s-1} + \gamma N(1 - H^{\frac{1}{N}})^{N-1}} dH$$

Changing variables $x = 1 - H^{1/N}$, we obtain

$$E[z_N] = c + \int_0^1 \frac{\mu_1(v-c)N(1-x)^{N-1}}{\sum_{s=1}^{S-1} s\mu_s x^{s-1} + \gamma N x^{N-1}} dx$$

We are interested in the sign of $E[z_{N+1}] - E[z_N]$. Then we can write:

$$\begin{aligned}
\frac{E[z_{N+1}] - E[z_N]}{\mu_1(v-c)} &= \int_0^1 \left(\frac{(N+1)(1-x)^N}{\sum_{s=1}^{S-1} s\mu_s x^{s-1} + \gamma(N+1)x^N} - \frac{N(1-x)^{N-1}}{\sum_{s=1}^{S-1} s\mu_s x^{s-1} + \gamma N x^{N-1}} \right) dx \\
&= \int_0^1 \frac{(N+1)(1-x)^N \left(\sum_{s=1}^{S-1} s\mu_s x^{s-1} + \gamma N x^{N-1}\right)}{\left(\sum_{s=1}^{S-1} s\mu_s x^{s-1} + \gamma(N+1)x^N\right)\left(\sum_{s=1}^{S-1} s\mu_s x^{s-1} + \gamma N x^{N-1}\right)} dx \\
&- \int_0^1 \frac{N(1-x)^{N-1} \left(\sum_{s=1}^{S-1} s\mu_s x^{s-1} + \gamma(N+1)x^N\right)}{\left(\sum_{s=1}^{S-1} s\mu_s x^{s-1} + \gamma(N+1)x^N\right)\left(\sum_{s=1}^{S-1} s\mu_s x^{s-1} + \gamma N x^{N-1}\right)} dx \\
&= \int_0^1 \frac{\left(\sum_{s=1}^{S-1} s\mu_s x^{s-1}\right)(1-x)^{N-1}(1-x(N+1))}{\left(\sum_{s=1}^{S-1} s\mu_s x^{s-1} + \gamma(N+1)x^N\right)\left(\sum_{s=1}^{S-1} s\mu_s x^{s-1} + \gamma N x^{N-1}\right)} dx \\
&+ \int_0^1 \frac{\gamma N(N+1)x^{N-1}(1-x)^{N-1}(1-2x)}{\left(\sum_{s=1}^{S-1} s\mu_s x^{s-1} + \gamma(N+1)x^N\right)\left(\sum_{s=1}^{S-1} s\mu_s x^{s-1} + \gamma N x^{N-1}\right)} dx
\end{aligned}$$

For convenience, let us refer to these two last integrals as I_2 and I_3, respectively. We now note that I_2 becomes arbitrarily close to zero when $\gamma \to 1$ (in that case $\sum_{s=1}^{S-1} s\mu_s x^{s-1} \to 0$). Let us now consider integral I_3. This integral takes on positive

values only when $x < 1/2$. Therefore we have:

$$I_3 = \int_0^{1/2} \frac{\gamma N(N+1)x^{N-1}(1-x)^{N-1}(1-2x)}{\left(\sum_{s=1}^{S-1} s\mu_s x^{s-1} + \gamma(N+1)x^N\right)\left(\sum_{s=1}^{S-1} s\mu_s x^{s-1} + \gamma N x^{N-1}\right)}dx$$
$$- \int_{\frac{1}{2}}^{1} \frac{\gamma N(N+1)x^{N-1}(1-x)^{N-1}(2x-1)}{\left(\sum_{s=1}^{S-1} s\mu_s x^{s-1} + \gamma(N+1)x^N\right)\left(\sum_{s=1}^{S-1} s\mu_s x^{s-1} + \gamma N x^{N-1}\right)}dx$$

Since the denominator of these integrals is monotonically increasing in x, we can write

$$I_2 > \int_0^{\frac{1}{N+1}} \frac{\gamma N(N+1)x^{N-1}(1-x)^{N-1}(1-2x)}{\left(\sum_{s=1}^{S-1} s\mu_s^{s-1}\left(\frac{1}{N+1}\right) + \gamma(N+1)\left(\frac{1}{N+1}\right)^N\right)\left(\sum_{s=1}^{S-1} s\mu_s^{s-1}\left(\frac{1}{N+1}\right) + \gamma N\left(\frac{1}{N+1}\right)^{N-1}\right)}dx$$
$$- \int_{\frac{1}{N+1}}^{1} \frac{\gamma N(N+1)x^{N-1}(1-x)^{N-1}(2x-1)}{\left(\sum_{s=1}^{S-1} s\mu_s\left(\frac{1}{N+1}\right)^{s-1} + \gamma(N+1)\left(\frac{1}{N+1}\right)^N\right)\left(\sum_{s=1}^{S-1} s\mu_s^{s-1}\left(\frac{1}{N+1}\right) + \gamma N\left(\frac{1}{N+1}\right)^{N-1}\right)}dx$$
$$= \frac{\int_0^1 \gamma N(N+1)x^{N-1}(1-x)^{N-1}(1-2x)dx}{\left(\sum_{s=1}^{S-1} s\mu_s^{s-1}\left(\frac{1}{N+1}\right) + \gamma(N+1)\left(\frac{1}{N+1}\right)^N\right)\left(\sum_{s=1}^{S-1} s\mu_s^{s-1}\left(\frac{1}{N+1}\right) + \gamma N\left(\frac{1}{N+1}\right)^{N-1}\right)} = 0.$$

Therefore we conclude that when γ is sufficiently large then $E[z_{N+1}] - E[z_N] > 0$. The critical $\widehat{\gamma}(N)$ is given by the solution to $I_2 + I_3 = 0$. Note that this is a sufficient condition and therefore not necessary. In fact, numerical simulations of the model show that this result holds for all γ. The proof of the Proposition is now complete. ■

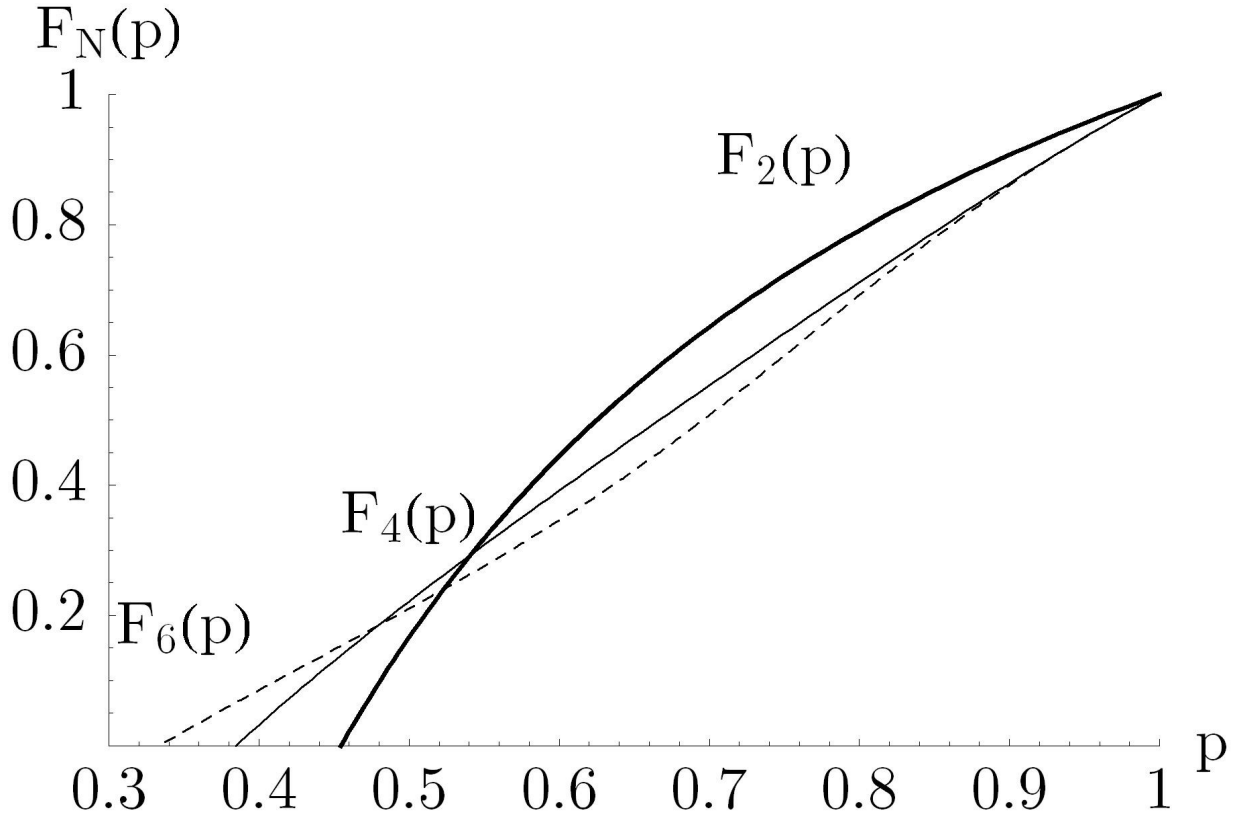

Figure 1: $F_N(p)$ for $N = 2, 4, 6$.

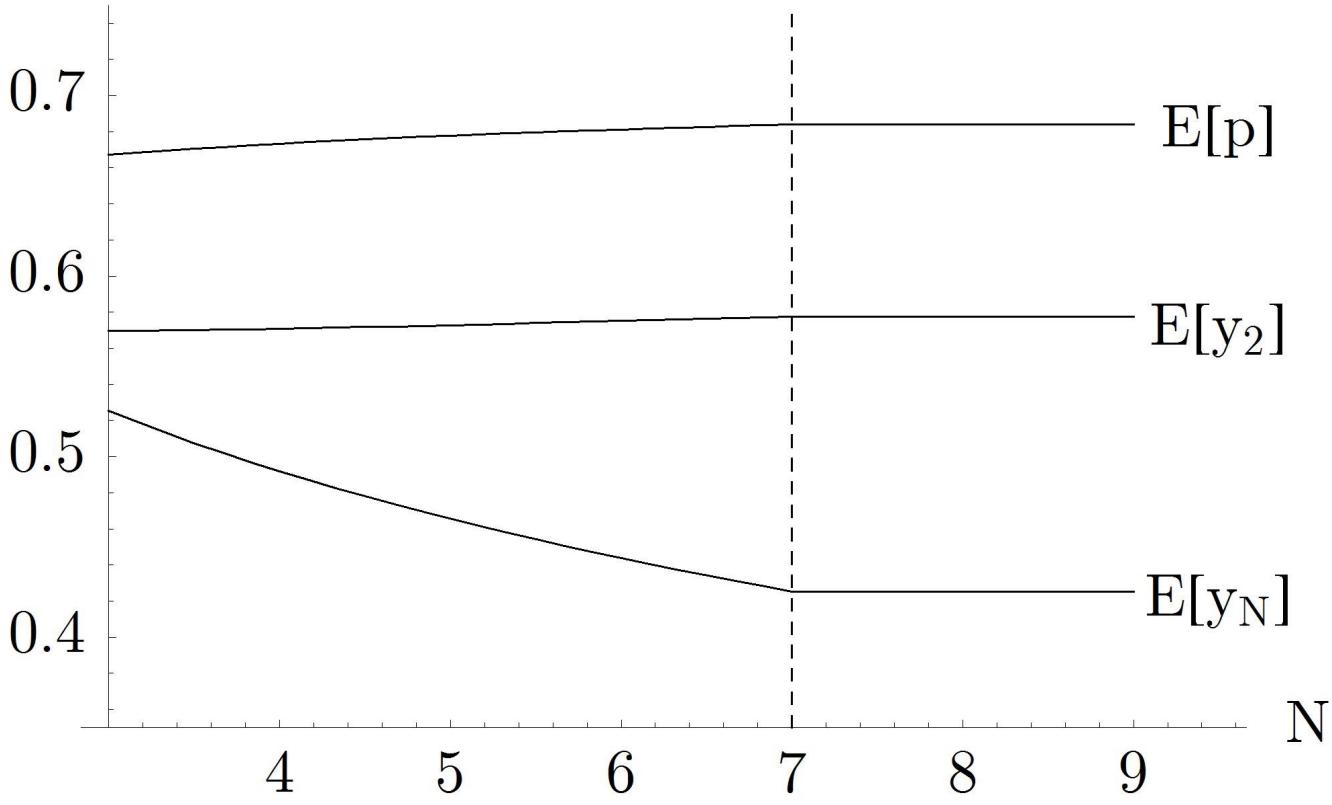

Figure 2: Expected price paid when observing s prices and the number of stores

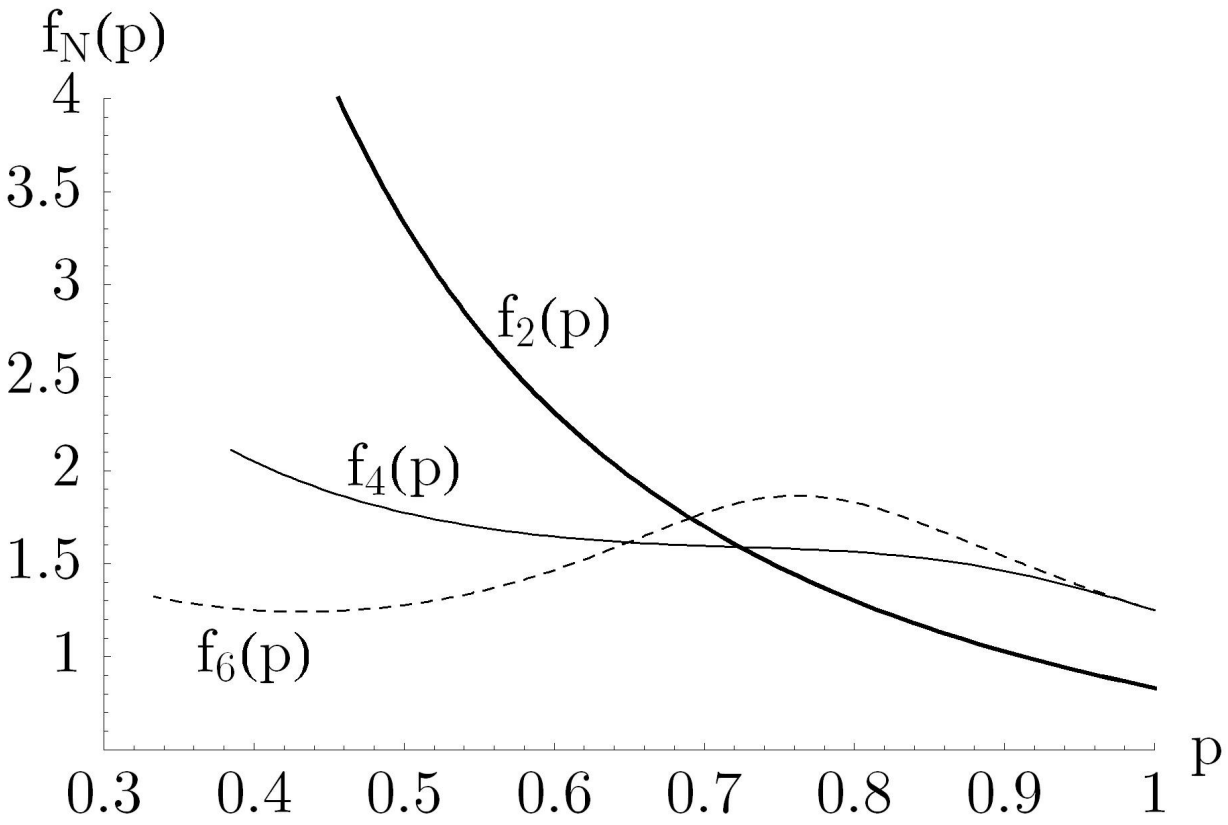

Figure 3: Density functions for $N = 2, 4, 6$.

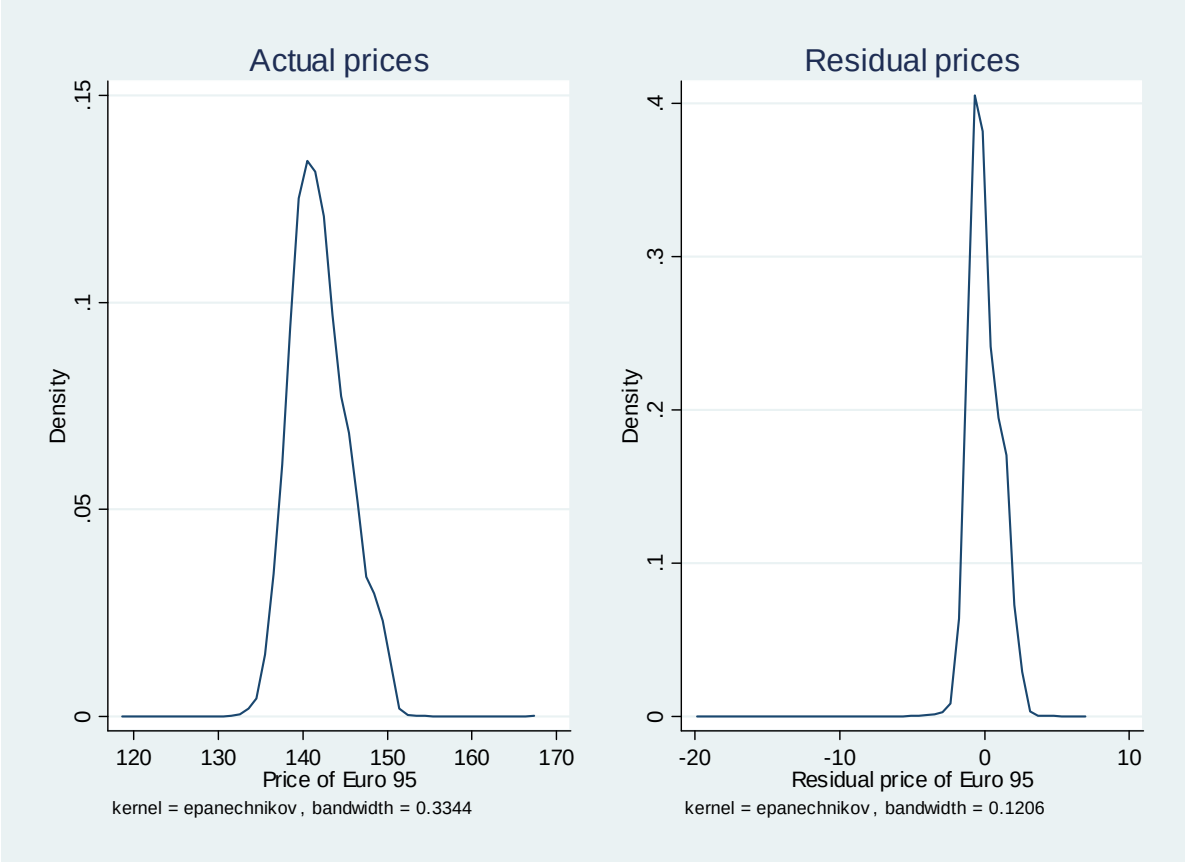

Figure 4: Density of Euro 95 raw and residual prices in the Netherlands

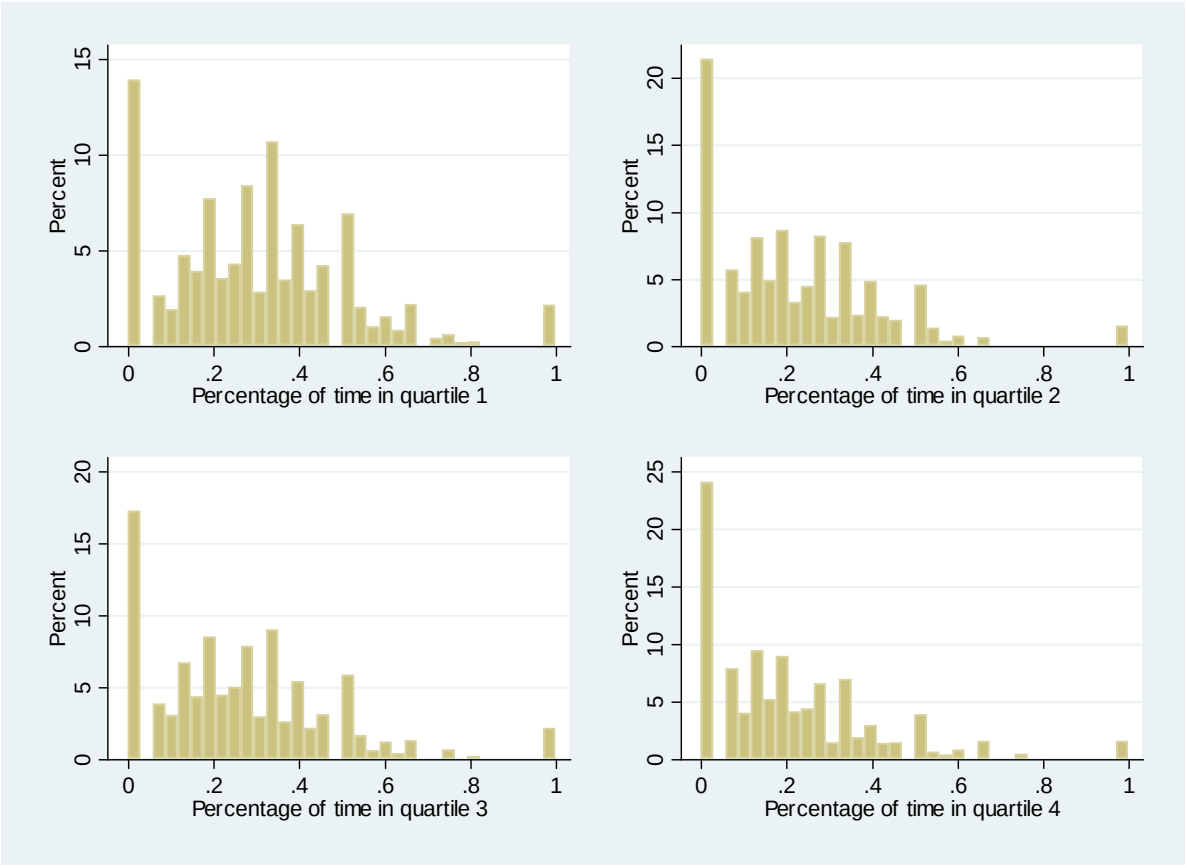

Figure 5: Time spent in each quartile of cross-sectional price distribution

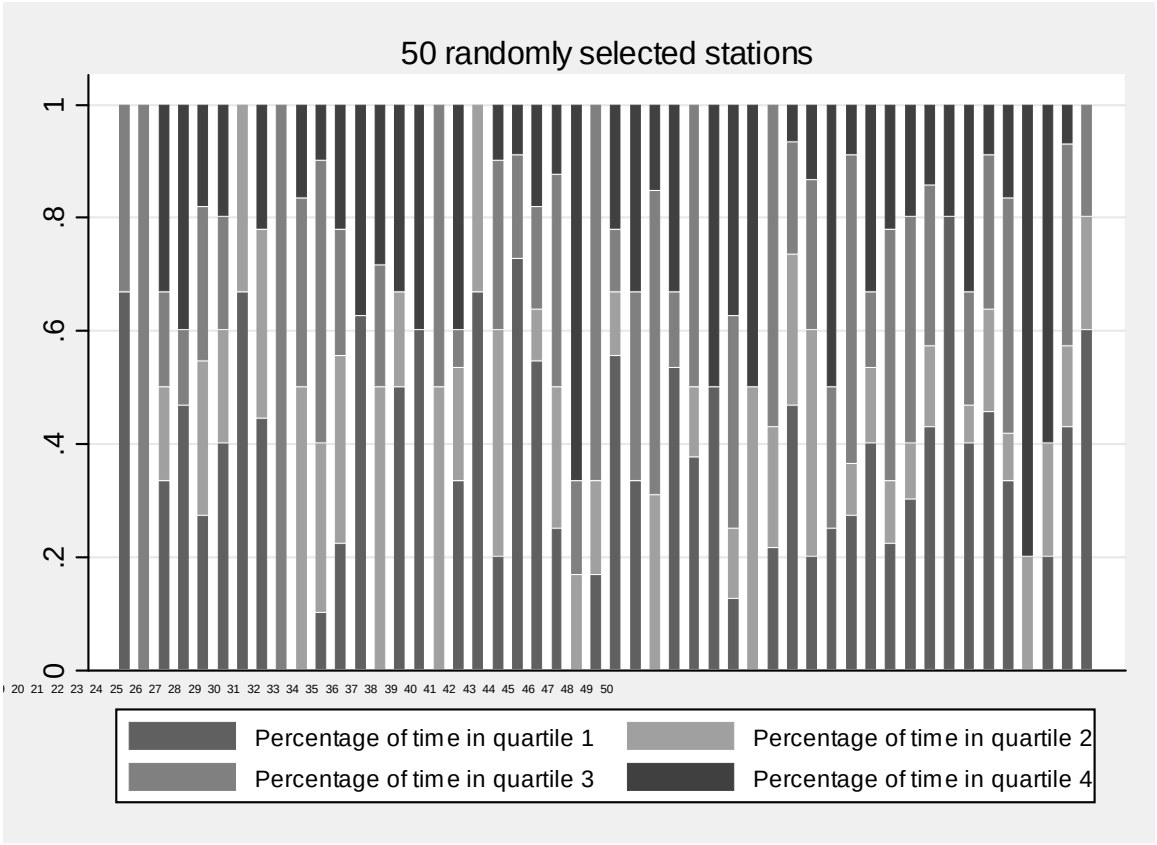

Figure 6: T1-T4 per gas station

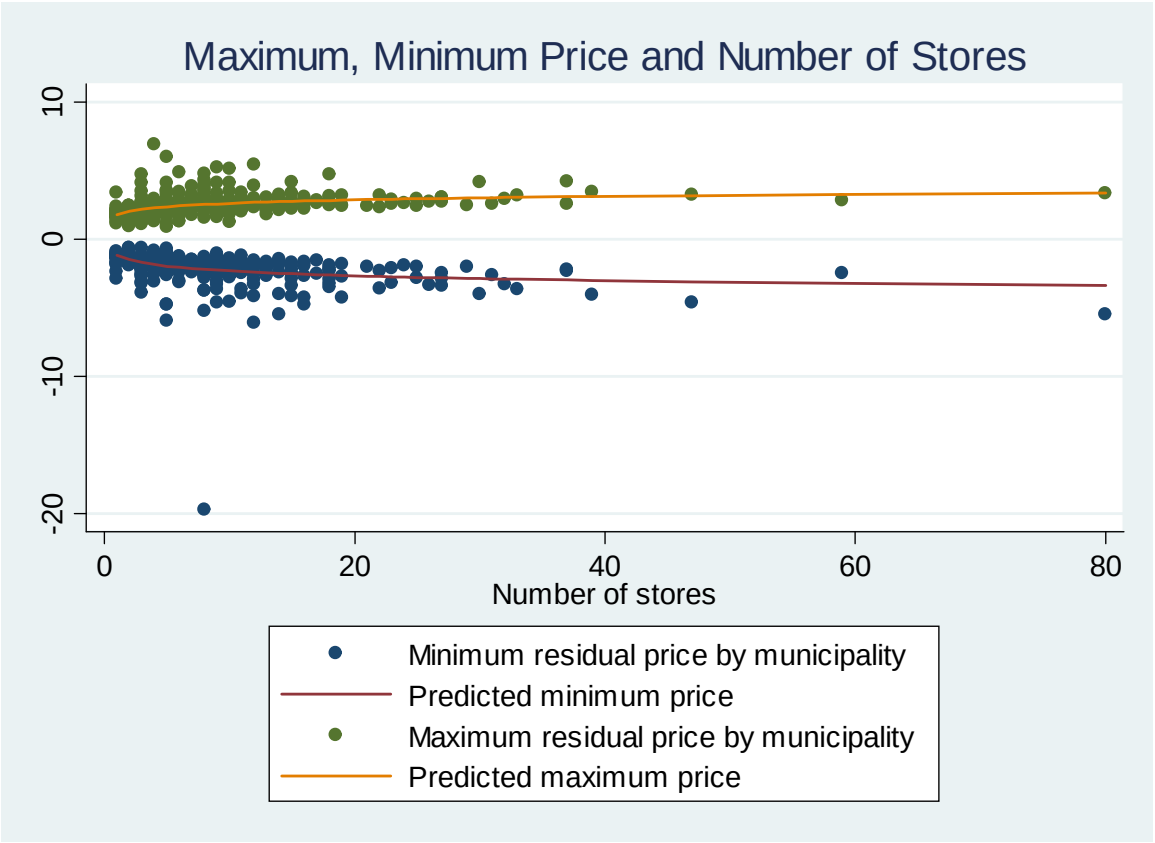

Figure 7: Extreme prices and the number of stores

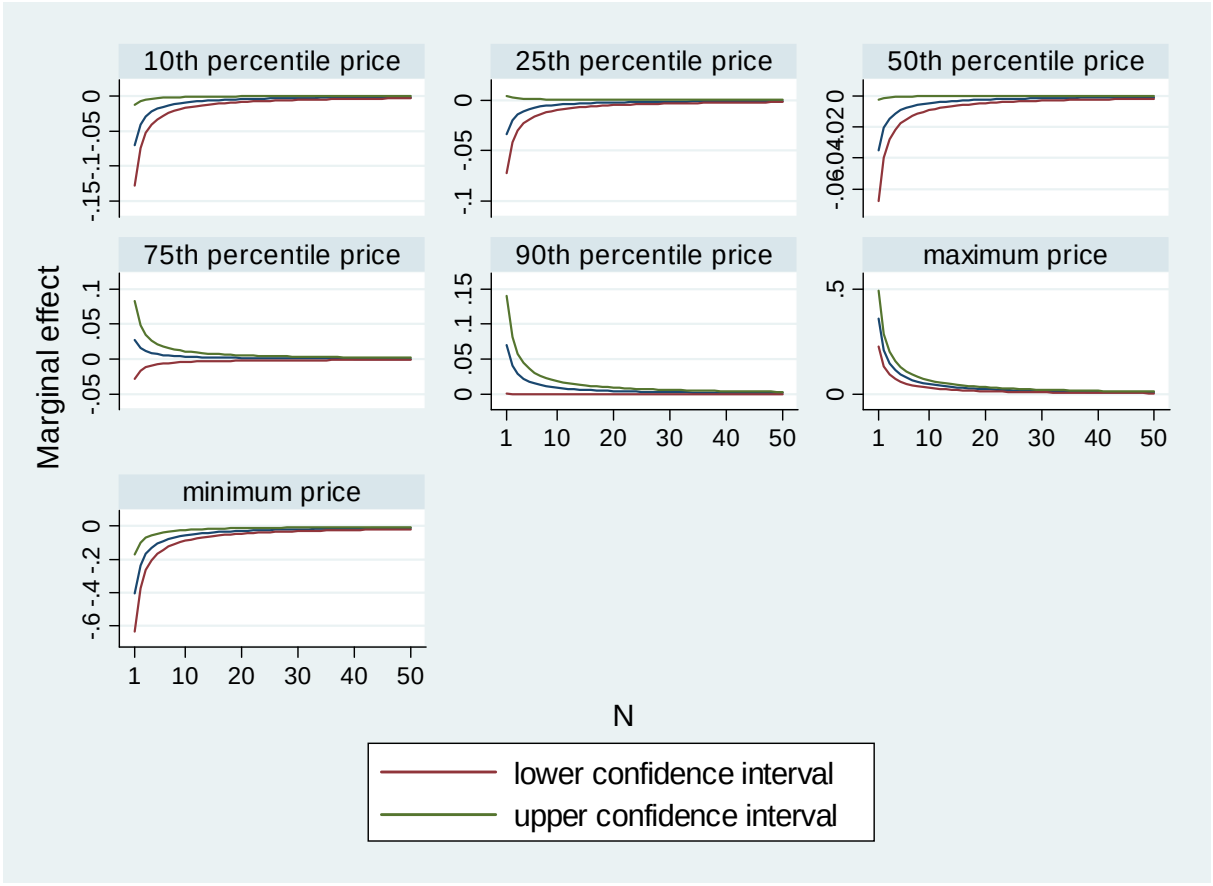

Figure 8: Change in price statistic and number of stores

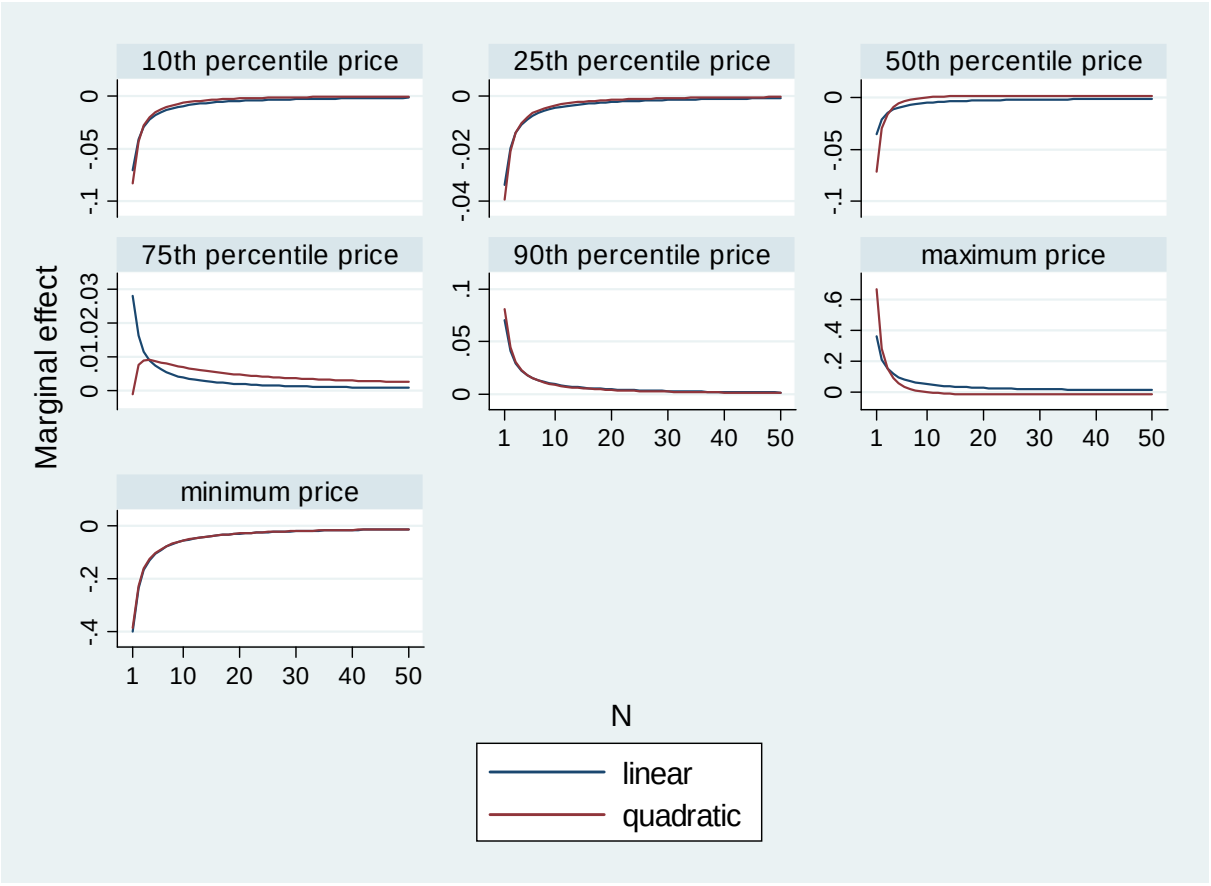

Figure 9: Marginal effects of squared and linear specifications

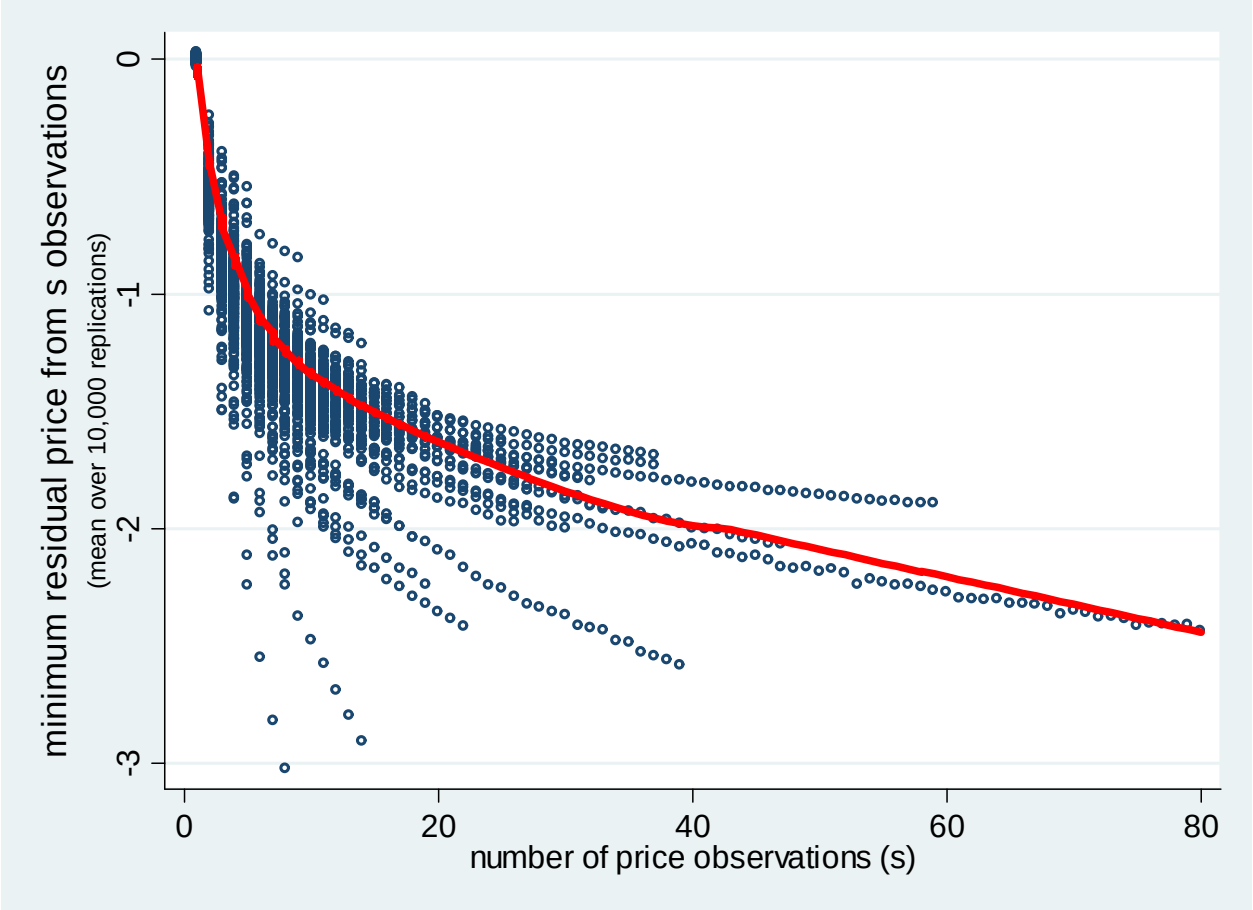

Figure 10: Expected price paid and number of price obsrevations

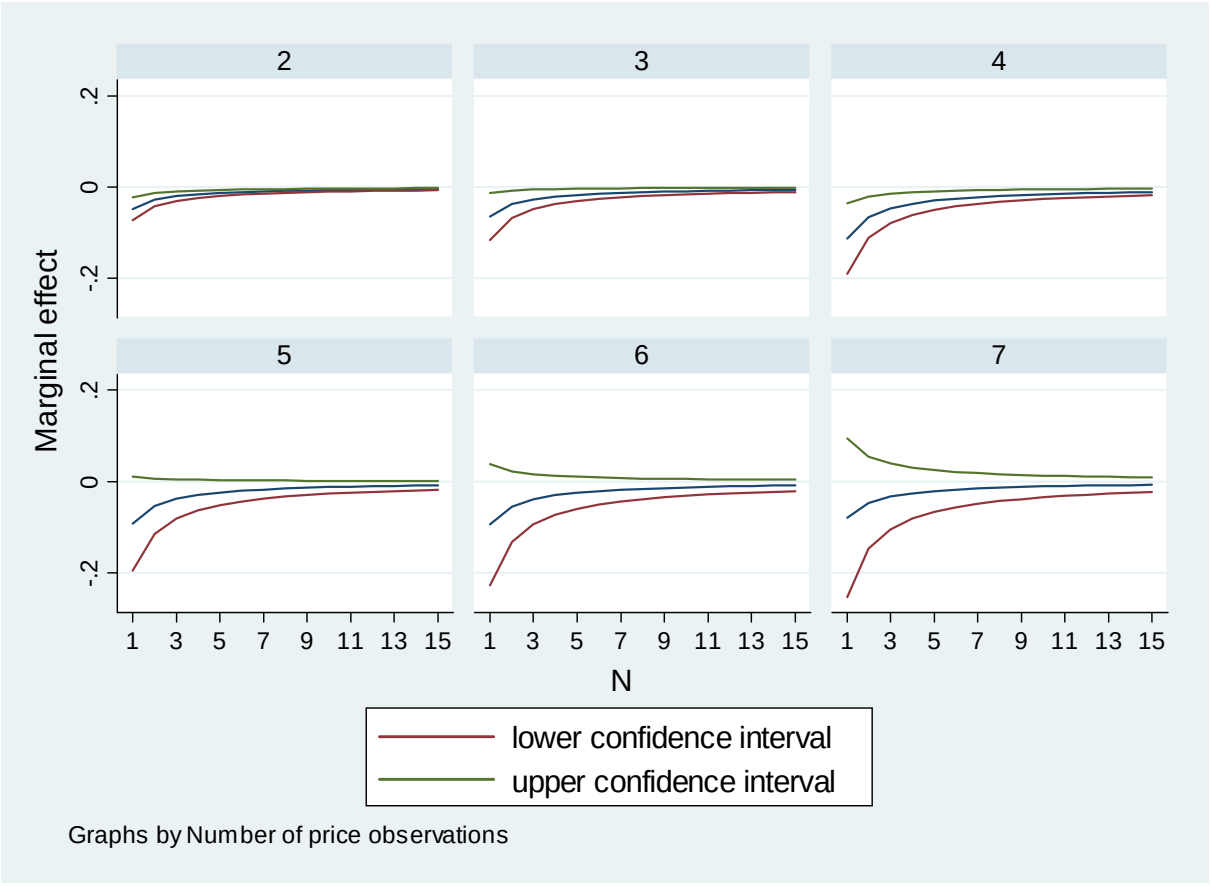

Figure 11: Change in expected price paid when observing s prices and number of stations

Table 1. Distribution of number of strores across markets

Number of stores	Frequency	Percent	Cumulative
1	16	3.6	3.6
2	35	8.0	11.6
3	45	10.2	21.8
4	48	10.9	32.7
5	56	12.7	45.5
6	36	8.2	53.6
7	26	5.9	59.6
8	30	6.8	66.4
9	24	5.5	71.8
10	21	4.8	76.6
11	19	4.3	80.9
12	17	3.9	84.8
13	8	1.8	86.6
14	10	2.3	88.9
15	8	1.8	90.7
16	6	1.4	92.1
17	2	0.5	92.5
18	7	1.6	94.1
19	3	0.7	94.8
21	1	0.2	95.0
22	2	0.5	95.5
23	2	0.5	95.9
24	1	0.2	96.1
25	2	0.5	96.6
26	1	0.2	96.8
27	3	0.7	97.5
29	1	0.2	97.7
30	1	0.2	98.0
31	1	0.2	98.2
32	1	0.2	98.4
33	1	0.2	98.6
37	2	0.5	99.1
39	1	0.2	99.3
47	1	0.2	99.6
59	1	0.2	99.8
80	1	0.2	100.0
Total	440	100	

Table 2. One-week transtition matrix (percentages)

		t+7			
		q_1	q_2	q_3	q_4
t	q_1	33	22	22	22
	q_2	27	27	27	18
	q_3	32	23	28	17
	q_4	27	22	31	20

A station enters the calulations only when it has data at t and at t+7. Entries in each are weighted averages of the week-specific transtition probabilities for each day t with weights equal to the share of observations in day t in the originating quartile out of the total number of observations for all days. Entries may not add up to 100 due to rounding.

Table 3. Residual price distribution by number of stores

Number of stores	Minimum price	10th percentile	25th percentile	Median price	75th percentile	90th percentile	Maximum price
1	-1.54	-1.23	-0.74	-0.13	0.81	1.4	1.88
2	-1.45	-1.12	-0.72	-0.17	0.64	1.39	1.92
3	-1.9	-1.22	-0.76	-0.16	0.8	1.48	2.25
4	-1.73	-1.18	-0.75	-0.17	0.75	1.47	2.38
5	-2.02	-1.21	-0.76	-0.16	0.7	1.49	2.44
6	-1.85	-1.19	-0.76	-0.19	0.72	1.53	2.38
7	-1.94	-1.23	-0.76	-0.19	0.74	1.54	2.51
8	-2.83	-1.26	-0.79	-0.18	0.79	1.54	2.77
9	-2.27	-1.23	-0.74	-0.22	0.74	1.52	2.8
10	-2.27	-1.23	-0.79	-0.19	0.74	1.52	2.65
11	-2.27	-1.22	-0.79	-0.18	0.81	1.51	2.59
12	-2.37	-1.23	-0.78	-0.18	0.75	1.53	2.75
13	-2.12	-1.15	-0.72	-0.16	0.68	1.44	2.29
14	-2.57	-1.24	-0.76	-0.16	0.72	1.59	2.65
15	-2.68	-1.24	-0.75	-0.15	0.73	1.44	2.83
16	-2.99	-1.16	-0.7	-0.15	0.71	1.41	2.59
17	-2.18	-1.32	-0.87	-0.2	0.9	1.58	2.67
18	-2.68	-1.24	-0.76	-0.18	0.74	1.54	3.11
19	-2.89	-1.25	-0.75	-0.21	0.81	1.49	2.64
21	-2.06	-1.2	-0.74	-0.27	0.74	1.49	2.41
22	-3.02	-1.36	-0.78	-0.1	0.88	1.47	2.73
23	-2.88	-1.16	-0.73	-0.14	0.67	1.43	2.76
24	-1.95	-1.31	-0.93	-0.18	0.86	1.79	2.61
25	-2.43	-1.19	-0.71	-0.2	0.68	1.45	2.64
26	-3.37	-1.32	-0.82	-0.19	0.8	1.54	2.72
27	-2.94	-1.25	-0.83	-0.21	0.75	1.53	2.88
29	-2.05	-1.34	-0.84	-0.14	0.71	1.4	2.47
30	-4.04	-1.24	-0.81	-0.19	0.85	1.58	4.16
31	-2.64	-1.24	-0.77	-0.18	0.72	1.46	2.53
32	-3.34	-1.18	-0.74	-0.19	0.8	1.37	2.89
33	-3.7	-1.17	-0.8	-0.17	0.52	1.45	3.14
37	-2.31	-1.22	-0.76	-0.26	0.75	1.55	3.4
39	-4.07	-1.28	-0.76	-0.09	0.76	1.64	3.43
47	-4.66	-1.26	-0.76	-0.23	0.73	1.61	3.23
59	-2.5	-1.25	-0.79	-0.18	0.78	1.45	2.8
80	-5.52	-1.28	-0.74	-0.25	0.75	1.54	3.32

Entries are weighted averages of each residual price statistic over all markets (municipalities) where the weights are the municipality's share of observations. Unweighted results are very similar.

Table 4. Effect of number of gas stations on the price distribution

	minimum price	10th percentile	25th percentile	median price	75th percentile	90th percentile	maximum price
				Panel A: OLS without controls			
Log (number of stations)	-0.508***	-0.0266	-0.0163*	-0.0161*	0.00715	0.0438**	0.368***
	(0.0523)	(0.0175)	(0.00958)	(0.00897)	(0.0138)	(0.0173)	(0.0341)
R-squared	0.11	0.01	0.01	0.01	0.00	0.01	0.16
				Panel B: OLS with 39 provincial dummies			
Log (number of stations)	-0.549***	-0.0285	-0.0209**	-0.0239**	0.0165	0.0513***	0.356***
	(0.0606)	(0.0194)	(0.0106)	(0.00960)	(0.0164)	(0.0196)	(0.0408)
R-squared	0.17	0.09	0.11	0.17	0.12	0.14	0.25
				Panel C: OLS with 39 provincial dummies and other controls			
Log (number of stations)	-0.305***	-0.0168	-0.0248	-0.0227	-0.000228	0.0428	0.320***
	(0.0886)	(0.0337)	(0.0196)	(0.0189)	(0.0312)	(0.0373)	(0.0640)
Tests:							
Other controls zero (p-value)	0.00	0.98	0.65	0.33	0.95	0.14	0.64
R-squared	0.19	0.09	0.13	0.20	0.13	0.16	0.27
				Panel D: 2SLS with 39 provincial dummies and other controls			
Log (number of stations)	-0.580***	-0.101**	-0.0487*	-0.0507**	0.0403	0.102**	0.518***
	(0.171)	(0.0423)	(0.0280)	(0.0241)	(0.0406)	(0.0510)	(0.0970)
Tests:							
Other controls zero (p-value)	0.37	0.77	0.44	0.13	0.95	0.04	0.46
J-test (p-value)	0.99	0.90	0.70	0.40	0.60	0.29	0.59
R-squared	0.18	0.07	0.13	0.20	0.12	0.16	0.26

Other controls include: average income per household, share of business cars, area (km^2), land area, urbanized and agrarian land shares, road length (km) and the number of sampled observations by market.

The top two panels are based on 439 observations. The bottom two panels are based on 423 observations; we loose 16 observations because of missing municipality-level data on the other controls. The instruments in the 2SLS panel are population size and local tax rates, both in logs.

Robust standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1

Table 5. Determinants of the Number of Stations

	Dep. var.: log(number of stations)			
	(1)	(2)	(3)	(4)
Log(population)	0.821***	0.805***	0.760***	0.758***
	(0.0255)	(0.0251)	(0.0633)	(0.0637)
Log(municipal tax)	-0.189***	-0.176***	-0.0850	-0.0811
	(0.0663)	(0.0668)	(0.0659)	(0.0666)
Average income per hh	--	--	0.00162	0.00121
			(0.0112)	(0.0112)
Share of business cars	--	--	1.515***	1.514***
			(0.543)	(0.543)
Area	--	--	0.0000664	0.0000700
			(0.000384)	(0.000386)
Land area	--	--	0.00365***	0.00366***
			(0.000710)	(0.000711)
Urbanized land share	--	--	-0.00284	-0.00286
			(0.00292)	(0.00292)
Agragrian land share	--	--	-0.00201	-0.00203
			(0.00165)	(0.00165)
Road length (km)	--	--	-0.00128***	-0.00128***
			(0.000265)	(0.000265)
Sample size	--	--	0.00300***	0.00301***
			(0.000695)	(0.000696)
F-test for significance of IV's	580.1	602.9	72.2	72.2
Provincial Effects	No	Yes	Yes	Yes
Observations	440	440	424	423
R-squared	0.72	0.81	0.84	0.84

Robust standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1

Table 6. Robustness checks I

	4 Size Groups: Two-stage 2SLS with 39 provincial dummies and other controls						
	minimum price	**10th percentile**	**25th percentile**	**median price**	**75th percentile**	**90th percentile**	**maximum price**
3 - 6 Stations	-0.4246***	-0.1440**	-0.0939**	-0.0629*	0.0801	0.2209***	0.6927***
	(.1649)	(.0671)	(.0435)	(.0366)	(.0619)	(.0792)	(.1223)
7 - 10 Stations	-0.4154	-0.0370	-0.0098	-0.0501**	0.0106	0.0179	0.2410**
	(.3464)	(.0378)	(.0265)	(.0234)	(.0388)	(.0519)	(.1146)
11+ Stations	-0.3321	-0.0036	-0.0296	0.0193	0.0617	0.0779	0.0355
	(.3229)	(.0404)	(.0279)	(.0233)	(.0396)	(.0569)	(.1584)
Tests:							
Other controls zero (p-value)	0.03	0.99	0.49	0.32	0.88	0.08	0.50
R-squared	0.19	0.08	0.12	0.20	0.13	0.14	0.25

Other controls include: average income per household, share of business cars, area (km^2), land area, urbanized and agrarian land shares, road length (km) and the number of sampled observations by market.

The instruments in the 2SLS regression are the predicted probabilities of belonging to a group size. The equation is just idenitifed. These predictions were obtained from a first-stage probit regression of each size dummy on population size and local tax rates (both in logs) as well as on the other controls and provincial dummies.

The number of observations in each regression is 423 municipalities. Robust standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1

	Raw prices: 2SLS with 39 provincial dummies and other controls							
	minimum price	**10th percentile**	**25th percentile**	**mean price**	**median price**	**75th percentile**	**90th percentile**	**maximum price**
Log (number of stations)	-1.423***	-1.434***	-0.818**	-0.184	-0.251	0.402	1.336***	1.354***
	(.376)	(.373)	(.387)	(.384)	(.421)	(.432)	(.415)	(.414)
Tests:								
Other controls zero (p-value)	0.00	0.00	0.02	0.05	0.02	0.07	0.01	0.00
J-test (p-value)	0.28	0.47	0.52	0.25	0.33	0.23	0.06	0.18
R-squared	0.47	0.45	0.41	0.44	0.42	0.38	0.38	0.41

Other controls include: average income per household, share of business cars, area (km^2), land area, urbanized and agrarian land shares, road length (km) and the number of sampled stations in each market-day. The instruments are population size and local tax rates, both in logs.

The number of market-day observations in each regression is 7091. Standard errors clustered at the municipality level in parentheses. *** p<0.01, ** p<0.05, * p<0.1

Table 7. Robustness checks II

	Adding Neighbours : 2SLS with 39 provincial dummies and other controls						
	minimum price	**10th percentile**	**25th percentile**	**median price**	**75th percentile**	**90th percentile**	**maximum price**
Log (number of stations)	-0.567***	-0.0950**	-0.0422	-0.0493**	0.0272	0.0886*	0.502***
	(0.176)	(0.0430)	(0.0292)	(0.0235)	(0.0414)	(0.0520)	(0.101)
Log (number of neighbouring stations)	0.0188	-0.00210	-0.00829	0.00587	0.0214	0.00466	0.00684
	(0.0773)	(0.0221)	(0.0158)	(0.0134)	(0.0223)	(0.0293)	(0.0493)
Tests:							
Other controls zero (p-value)	0.42	0.45	0.55	0.44	0.81	0.12	0.48
J-test (p-value)	0.99	0.90	0.70	0.40	0.60	0.29	0.59
R-squared	0.18	0.08	0.13	0.14	0.14	0.15	0.25

Other controls include: average income per household, share of business cars, area (km^2), land area, urbanized and agrarian land shares, road length (km) and the number of sampled observations by market. The instruments are population size and local tax rates, both in logs.
Based on 420 observations corresponding to municipalities with non-zero number of neighbours.
Robust standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1

	Diesel: 2SLS with 39 provincial dummies and other controls						
	minimum price	**10th percentile**	**25th percentile**	**median price**	**75th percentile**	**90th percentile**	**maximum price**
Log (number of stations)	-0.474**	-0.0791	-0.0736***	-0.0146	0.0738**	0.0910**	0.502***
	(0.214)	(0.0553)	(0.0274)	(0.0202)	(0.0317)	(0.0462)	(0.101)
Tests:							
Other controls zero (p-value)	0.20	0.16	0.18	0.23	0.41	0.44	0.21
J-test (p-value)	0.81	0.15	0.45	0.06	0.13	0.41	0.17
R-squared	0.16	0.16	0.17	0.10	0.14	0.12	0.29

Other controls include: average income per household, share of business cars, area (km^2), land area, urbanized and agrarian land shares, road length (km) and the number of sampled observations by market. The instruments are population size and local tax rates, both in logs.
Based on 424 observations. Robust standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1

Table 8. Expected price paid and the number of stations

	Dep. Var: Expected price paid when observing s prices, $E[Min\{p_1,p_2,...,p_s\}]$					
Number of observations (s)	**s = 2**	**s = 3**	**s = 4**	**s = 5**	**s = 6**	**s = 7**
Log (number of stations)	-0.0689***	-0.0924**	-0.163***	-0.133*	-0.136	-0.115
	(0.0182)	(0.0375)	(0.0564)	(0.0760)	(0.0975)	(0.127)
Tests:						
Other controls zero (p-value)	0.08	0.15	0.15	0.14	0.22	0.15
J-test (p-value)	0.93	0.62	0.59	0.78	0.27	0.26
Number of observations	407	373	329	282	227	192
R-squared	0.041	0.012	0.007	0.021	0.058	0.068

Other controls include: average income per household, share of business cars, area (km^2), land area, urbanized and agrarian land shares, road length (km) and the number of sampled observations by market.
The instruments are population size and local tax rates, both in logs.
Robust standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1