Aggregate Housing Price Trends in Chicago and Local Characteristics

Emiliano Harris

Submitted in partial fulfillment of the
requirements for the degree of
Master of Science at the
École Nationale de Statistique et d'Administration Économique

October 2017

Aggregate Housing Price Trends in Chicago and Local Characteristics

ENSAE Thesis

Emiliano Harris*

October 2017

Contents

*I would like to thank Amine Ouazad for advising me on this work and kindly introducing me to new topics in urban economics.

1 Introduction

Accurately tracking housing price trends is a methodologically complex endeavor. Indeed, relying solely on year-to-year price means, for instance, would not provide a satisfactory estimate, because houses are a differentiated good, and thus the characteristics of houses which are sold vary from one year to the next. Consequently the evolution captured by an average (or by a median, for that matter) may be entirely due to the differences in characteristics of the houses sold in various years. A typical strategy is to rely on hedonic regressions, for instance by regressing prices on as long a list of house characteristics as available, as well as time dummies. The time dummy coefficient estimates are then price indices. This specific method has the disadvantage of constraining the implicit prices of characteristics to be constant over time, but overall, the main drawback of hedonic method is the richness of the data that it requires. It could always be assumed that some unobserved characteristic plays a fundamental role in house differentiation, thus leading to incorrect estimates.

Bailey, Muth and Nourse (1963) introduced the repeat sales methodology to remedy this problem. The central idea is to restrict the data set of house sales to houses which have been sold at least twice. Their method enables a valid comparison of house prices from one year to the next, by keeping characteristics constant. This method was improved later on, mainly by Case and Shiller (1987), who argued that heteroscedastic shocks on individual house prices should be taken into account. In this thesis, I review repeat sales indices and introduce a method to address sample selection bias by adapting Heckman's (1979) approach to these indices.

Addressing this sample selection bias is crucial because despite the major advance represented by the repeat sales method, it imposes a very demanding condition on the data set, namely that the houses be sold at least twice during the period of observation. This may potentially result in a very distinctive set of houses, since – as shown in a further section – houses that transact are located in particular types of neighborhoods. In a nutshell, the repeat sales indices may be subject to sample selection issues, and may hence be biased. My main focus during the internship was thus to correct the repeat sales index for potential sample selection, by taking advantage of a very rich data set.

2 Modeling housing price dynamics: the repeat sales approach

This section is devoted to the presentation of the repeat sales method, and its developments by Case and Shiller (1987, 1991). Bailey, Muth and Nourse assumed that

between two sales observations of a given house, characteristics remain constant. This is not realistic, because changes are likely to occur between two transactions, especially if they are separated by a long time interval. Case and Shiller (1987) reject this assumption, and propose a solution to put less weight on houses which have been sold after longer time intervals. Shiller (1991) goes on to distinguish the estimator from 1987, based on a geometric mean approach, from an estimator based on an arithmetic mean approach. The latter was adopted by real estate agencies and financial information firms which publish monthly housing indices, such as Standard & Poor's. Thus, to ease comparison between my index computations and publicly available indices, sections (3) and the following will exclusively focus on the arithmetic approach.

In section (2.1), I present the general model. Sections (2.2) and (2.3) introduce the Case-Shiller geometric repeat sales model (GRS) and the arithmetic repeat sales model (ARS), respectively.

2.1 Model of housing price evolution

Consider house $i \in I$, in period $t \in [\![0, T]\!]$. The repeat sales approach is based on the following equation of house price formation since its inception:

$$p_{it} = p_{at} + e_{it} + n_{it} \tag{1}$$

where p_{it} denotes the log price of house i in period t and p_{at} is the aggregate real estate log price index in period t. In Baily, Muth and Nourse's version, the only error term was n_{it}, a sale-specific random error with variance σ_n^2. Case and Shiller added e_{it}, a property-specific Gaussian random walk, hence the sum of previous steps $(\varepsilon_{is})_{s=0\dots t}$, each assumed to follow a normal distribution with mean 0 and variance σ_ε^2. Both n_{it} and e_{it} are assumed i.i.d., cross-sectionally and over time. The random walk represents potential changes occurring in a property's characteristics from one sale to another, or the impact of changes in taste on the property' value, as is made more clear in equations (2) and (3).

Equation (1) naturally leads to an expression for the price change of a given house, from one sale to the next. Indeed, the log ratio of prices for house i sold in period t_i and subsequently in t_i' ($t_i' > t_i$), is

$$p_{it_i'} - p_{it_i} = p_{at_i'} - p_{at_i} + \sum_{k=t_i+1}^{t_i'} \varepsilon_{ik} + n_{it_i'} - n_{it_i} \tag{2}$$

Importantly, note that the variance of the errors is time-dependent (one of Case and Shiller's contributions is to have accounted for this), and thus, this model

allows for heteroscedasticity:

$$\mathbb{V}(\sum_{k=t_i+1}^{t'_i} \varepsilon_{ik} + n_{it'_i} - n_{it_i}) = (t'_i - t_i)\sigma_\varepsilon^2 + 2\sigma_n^2 \tag{3}$$

This general framework indirectly defines an aggregate sales price. The purpose of the estimators is to recover it, and *ipso facto* to generate a housing price index.

2.2 The Geometric repeat sales estimator

The GRS is none other than the estimator resulting directly from equation (2): the aggregate sales prices are the parameters of interest, as is illustrated in the following regression equation

$$p_{it'_i} - p_{it_i} = \sum_{t=1}^{T} (\mathbb{1}\{t'_i = t\} - \mathbb{1}\{t_i = t\})p_{at} + e_{it'_i t_i} + n_{it'_i t_i} \tag{4}$$

where $e_{it'_i t_i}$ and $n_{it'_i t_i}$ are shorthand for the error terms from equation (2). Corresponding matrices are constructed as suggested by the above equation. Suppose $|I| = n$, and each house in I was sold exactly twice during the period of observation. Let $\mathbf{\Delta P}$ denote a vector of length n where entry i is $p_{it'_i} - p_{it_i}$. Let $\mathbf{Z}$ be an $n \times T$ matrix of regressors, where entry $z_{ij} = \mathbb{1}\{t'_i = j\} - \mathbb{1}\{t_i = j\}$; in other words, entry z_{ij} is equal to 1 if house i was sold for the second time in period j, -1 if it was sold for the first time in period j, and 0 if it wasn't sold in period j. Then, equation (4) can be rewritten

$$\mathbf{\Delta P} = \mathbf{Z}\beta + \boldsymbol{\iota} \tag{5}$$

where β denotes a vector of length T, which contains the parameters of interest, *i.e.* entry t is p_{at}, and $\boldsymbol{\iota}$ is an error vector of length n (ignoring for now the specific decomposition of errors). Then, assuming that individual error terms are uncorrelated, that they are homoscedastic and with mean 0, the Gauss-Markov theorem applies to the ordinary least squares estimator $\hat{\beta}_{\text{GRS}} = (\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{\Delta P}$.

It is worthwhile to consider the normal equations of this regression: $\mathbf{Z}'\mathbf{\Delta P} = \mathbf{Z}'\mathbf{Z}\hat{\beta}_{\text{GRS}}$. Suppose $n = 3$ and $T = 2$, such that the setting is as described in table (1).
The two normal equations resulting from this setting are

$$\hat{\beta}_{\text{GRS1}} = \hat{p}_{a1} = \frac{p_{2,1} + p_{3,1}}{2} - \frac{(p_{2,2} - \hat{p}_{a2}) + p_{3,0}}{2}$$
$$\hat{\beta}_{\text{GRS1}} = \hat{p}_{a2} = \frac{p_{1,2} + p_{2,2}}{2} - \frac{p_{1,0} + (p_{2,1} - \hat{p}_{a1})}{2}$$

Table 1: Setting of housing sales in a 3 periods and 3 houses-example

$i \backslash t$	0	1	2
1	$p_{1,0}$	$\diagdown$	$p_{1,2}$
2	$\diagdown$	$p_{2,1}$	$p_{2,2}$
3	$p_{3,0}$	$p_{3,1}$	$\diagdown$

The GRS index for period 1, $\hat{p}_{a1}$, is the difference between the log price average of all houses sold in period 1 (here houses 2 and 3), and the log price average of the same houses sold in the base period (*i.e.* period 0). Crucially, since house 2's other sale was not in the base period, but in period 2, its base-period price is inferred by subtracting $\hat{p}_{a2}$ from $p_{2,2}$. Similar observations can be drawn from the formula for $\hat{p}_{a2}$. Taking the exponential of these estimates gives geometric averages, thus justifying the name of this estimator.

Now suppose that the vector of errors is in fact the sum of two error vectors, in accordance with the general model of price evolution: $\boldsymbol{\iota} = \mathbf{e} + \mathbf{n}$. A version of feasible generalized least squares, proposed by Case and Shiller (1987), can be applied in this context, so as to take heteroscedasticity into account. Indeed, let $\boldsymbol{\tau}$ denote a vector of length n, where entry i is $t'_i - t_i$, and let $\hat{\boldsymbol{\iota}}^2$ be the vector of squared residuals from the OLS regression leading to $\hat{\beta}_{\text{GRS}}$. Then add two steps to the previous regression. First regress $\hat{\boldsymbol{\iota}}^2$ on $\boldsymbol{\tau}$ and $\mathbf{1}$, a vector with 1 in every entry: this provides an estimate $\hat{\sigma}^2_\varepsilon$ for σ^2_ε, and $2\hat{\sigma}^2_n$ for $2\sigma^2_n$, according to the variance equation (3) ; second, perform a weighted least squares version of the initial regression, or in other words, the initial regression with every term divided by the square root of the fitted values obtained in the second step:

$$\frac{p_{it'_i} - p_{it_i}}{\sqrt{(t'_i - t_i)\hat{\sigma}^2_\varepsilon + 2\hat{\sigma}^2_n}} = \frac{\sum_{t=1}^{T}(\mathbb{1}\{t'_i = t\} - \mathbb{1}\{t_i = t\})p_{at}}{\sqrt{(t'_i - t_i)\hat{\sigma}^2_\varepsilon + 2\hat{\sigma}^2_n}} + \frac{\iota_{it_it'_i}}{\sqrt{(t'_i - t_i)\hat{\sigma}^2_\varepsilon + 2\hat{\sigma}^2_n}}$$

2.3 The Arithmetic repeat sales estimator

By analogy with the GRS, the ARS is a means to obtain a comparable aggregate housing price estimate, in which the log prices are replaced by their levels, and the estimator modified in order to still control for the change in mix of houses across time.[1] Shiller argues that the ARS index has desirable properties, such as being a value-weighted index, which is important if price changes differ with house values,

[1] Simply replacing the log prices with their levels, while reproducing identically the GRS method would be equivalent to taking absolute differences in prices, instead of percentage changes as in the GRS, thus leading to overestimating price increases in years when expensive houses are sold.

and facilitating comparison and covariance analyses between housing portfolios and other assets, which are generally also based on arithmetic means.

The suggested regression equation for the ARS method should be

$$\mathbb{1}\{t_i = 0\}P_{it_i} = \sum_{t=1}^{T}(\mathbb{1}\{t'_i = t\}P_{it'_i} - \mathbb{1}\{t_i = t\}P_{it_i})p_{at}^{-1} + e_{it'_it_i} + n_{it'_it_i} \tag{6}$$

where P_{it} is the level price of house i in period t. As in the GRS estimator, $t = 0$ is the base year, which is why the index p_{at} is only expressed for $t > 0$ in the equation. The normal equations will clarify the presence of the index's reciprocal instead of the index itself.

However, prices, which are now independent variables, are stochastic variables, which depend for instance on the buyers' and sellers' imperfect appreciation of housing market price levels on the day of the sale. Hence restricting our method to regression equation (6) would result in an error in variables problem. One solution is to rely on the independent variable defined in the GRS equation (4) as an instrument. Consequently, the ARS estimator can be obtained *via* a two-stage least squares procedure, the first stage equation being

$$\sum_{t=1}^{T}\mathbb{1}\{t'_i = t\}P_{it'_i} - \mathbb{1}\{t_i = t\}P_{it_i} = \sum_{t=1}^{T}(\mathbb{1}\{t'_i = t\} - \mathbb{1}\{t_i = t\})\gamma_t + u_i \tag{7}$$

where u_i is an error term, assumed to be uncorrelated with the instrument, and uncorrelated across houses. Let $\mathbf{X}$ be the $n \times T$ matrix of regressands in the above equation. We have that entry $x_{ij} = z_{ij}P_{ij}$ (where z_{ij} is an element of matrix $\mathbf{Z}$ as defined in subsection 2.2): P_{ij} if house i was sold in period j for the second time, $-P_{ij}$ if it was sold in period j for the first time, and 0 otherwise. Furthermore, let $\mathbf{y}$ denote a vector of length n, where entry i is $\mathbb{1}\{t_i = 0\}P_{it_i}$. Hence, in matrix notation, regression equations (6) and (7) become respectively

$$\mathbf{y} = \mathbf{X}\beta + \boldsymbol{\iota} \tag{8}$$

$$\mathbf{X} = \mathbf{Z}\gamma + \mathbf{u} \tag{9}$$

where γ is a $T \times T$ matrix, in which each column is the vector $[\gamma_1, \ldots, \gamma_T]\top$, and $\mathbf{u}$ is a vector of errors. Since $\mathbf{X}$ and $\mathbf{Z}$ have the same dimensions, the ARS estimator is just identified, and can thus be expressed as an IV estimator: $\hat{\beta}_{\text{ARS}} = (\mathbf{Z}'\mathbf{X})^{-1}\mathbf{Z}'\mathbf{y}$. If $p\lim \mathbf{Z}'\mathbf{u}/n = 0$ and $p\lim \mathbf{Z}'\mathbf{X}/n$ is non-singular, then this estimator is consistent.

Consider again the example from table (1). This time, the normal equations, $\mathbf{Z}'\mathbf{y} = \mathbf{Z}'\mathbf{X}\hat{\beta}_{\text{ARS}}$, are

$$\hat{\beta}_{\text{ARS1}}^{-1} = \hat{p}_{a1} = \frac{P_{2,1} + P_{3,1}}{P_{2,2}\hat{p}_{a2}^{-1} + P_{3,0}}$$
$$\hat{\beta}_{\text{ARS2}}^{-1} = \hat{p}_{a2} = \frac{P_{1,2} + P_{2,2}}{P_{1,0} + P_{2,1}\hat{p}_{a1}^{-1}}$$

This estimator is arithmetic, because its expression for period t is the ratio of price averages for the houses sold in t, to their – potentially corrected – price average in the base period (to see this, divide both the numerator and denominator by the number of houses sold in t, in the right hand side of the normal equations).

Just as for the GRS estimator, heteroscedasticity can be corrected with a weighted least squares procedure, where the weights are the square root of the fitted values resulting from the regression of squared residuals $\hat{\boldsymbol{\iota}}^2$ on a vector of ones and time spells between two sales.

3 Accounting for selection in the repeat sales approach

The repeat sales approach provides methods to capture housing price variations while accounting for the intrinsic heterogeneity of houses, by convincingly keeping characteristics constant when prices are compared. However, it may be safely objected – and evidence for this is given in section (4) – that house heterogeneity not only impacts price differences and variations. It also is a key determinant of house sale. The repeat sales approach does not address this issue, and requires houses in the sample to have been sold at least twice during the period of observation. This requirement is a potential source of sample selection, or more specifically incidental truncation: it is conceivable that for a given house, two sales take place during the period of observation, only if some other variables take on a precise range of values. In particular, these variables can be expected to determine (i) the decision to sell, and (ii) the likelihood of finding a buyer. Put differently, the combination of self-selection into the housing market, and of housing demand, are likely to introduce non-randomness in the subset of houses which were sold at least twice, thus biasing estimates of housing price indices.

I seek to tackle sample selection by relying on a traditional solution to this issue: Heckman's correction method.

3.1 Modeling sample selection

We now suppose that being part of the subset of houses which were sold at least twice, denoted $\mathcal{S}$, can be modeled by a latent variable μ°, with $\mu_i^\circ \geq 0$ if and only if house i is in the selection sample. Otherwise, $\mu_i^\circ < 0$ if house i was not sold at least twice during the period of observation, in which case it belongs to set $\bar{\mathcal{S}}$. Since μ_i° is latent, we only observe $\mu_i = \mathbb{1}\{\mu_i^\circ \geq 0\}$, and make the assumption that

$$\mu_i^\circ = \mathbf{n}_i\delta + v_i \tag{10}$$

where $\mathbf{n}_i$ is a vector of explanatory variables for the presence or absence of house i in set $\mathcal{S}$, and v_i is an i.i.d. error term with a standard normal distribution. For simplicity, let $P_{it_it_i'}$ denote the pair of prices for house i. If $i \in \mathcal{S}$, then $P_{it_it_i'} = P^\circ_{it_it_i'} = (P_{it_i}, P_{it_i'})$, otherwise it is not observable. This leads to the setup described in table (2).

Table 2: Observed, latent and unobserved variables

	$\mu_i^\circ \geq 0$	$\mu_i^\circ < 0$
i	$\in \mathcal{S}$	$\in \bar{\mathcal{S}}$
μ_i	1	0
$P_{it_it_i'}$	$P^\circ_{it_it_i'}$	unobserved

I set vector notation based on the matrix notation from the ARS equations (8) and (9). Let y_i denote entry i of vector $\mathbf{y}$, $\mathbf{x}_i$ is a length T vector equal to the ith row of matrix $\mathbf{X}$, ι_i is entry i of $\boldsymbol{\iota}$, and $\mathbf{z}_i$ is a length T vector equal to the ith row of matrix $\mathbf{Z}$. We now have the following three-equation model:

$$y_i = \mathbf{x_i}\beta + \iota_i \tag{11}$$
$$\mathbf{x}_i = \mathbf{z}_i\gamma + u_i \tag{12}$$
$$\mu_i = \mathbb{1}\{\mathbf{n}_i\delta + v_i \geq 0\} \tag{13}$$

Note that these hypotheses do not, by themselves, necessarily imply that sample selection in $\mathcal{S}$ biases the estimation of the ARS index. Indeed, if $\mathbf{y}$ and the selection variable, μ°, are independent conditional on $\mathbf{X}$, then selection would be exogenous, and the ARS estimation would be unbiased. If we assume, however, that this is not the case, we may model selection bias as arising from a correlation between $\boldsymbol{\iota}$ and v_i. Thus, we consider the classic hypothesis of bivariate normality between the error terms:

$$\begin{bmatrix} \boldsymbol{\iota}_i \\ v_i \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma^2_{\iota_i} & \rho\sigma_{\iota_i} \\ \rho\sigma_{\iota_i} & 1 \end{bmatrix}\right) \tag{14}$$

where $\iota_i = e_{it'_i t_i} + n_{it'_i t_i}$, $\sigma^2_{\iota_i} = \mathbb{V}(\iota_i)$ (the full expression is given in equation (3)), and ρ is the correlation between ι_i and v_i.

Ideally, the selection equation should incorporate time-dependent selection variables, since local characteristics determining the selection of a given house into $\mathcal{S}$ are likely to evolve across time. However, the model would require some refinement, as it is not straightforward to adapt this requirement to houses in $\bar{\mathcal{S}}$; indeed, these houses are never sold during the period of observation, and thus no specific date can be attributed to them.

One could interpret as follows the bias that sample selection might lead to in the ARS estimation. Suppose for simplicity that all houses in the sample were sold in $t = 0$ and $t = 1$. Suppose furthermore that $p_{a1} \in (0, 1)$, and that the houses which were sold in period 1, *i.e.* houses in $\mathcal{S}$, all had second period prices such that $P_1 > \tilde{P}_1$, for some $\tilde{P}_1 > 0$. In words, house values decreased between the two periods, and the only houses which were sold in period 1, were houses whose price in the base period was above a threshold $\tilde{P}_1$, for instance because $\tilde{P}_1/p_{a1} \equiv \tilde{P}_0$ represents a minimum, expressed in base period-value, required by house owners willing to sell their house in period 1. In this setup, the price formation equation – using price levels instead of log prices and subsuming the error terms under the term ι_i –, based on equation (2), would be $P_{i0} = P_{i1}/(p_{a1}\iota_i)$.[2] House i, on average, has $P_{i0} = P_{i1}/p_{a1}$. However, some houses may experience individual shocks, such that $\iota_i \neq 1$. In particular, there may exist some house $i \in \mathcal{S}$ such that $P_{i0} < \tilde{P}_0$, but with $\iota_i > 1$ large enough so that $P_{i1} = P_{i0}p_{a1}\iota_i > \tilde{P}_1$. In other words, if it were to follow the aggregate house value trend, the price of such a house i in the base period would be too low to reach the minimum required for house i to be in $\mathcal{S}$, but because of an individual shock, its price in period 1 is above the threshold $\tilde{P}_1$. This possibility will induce a negative correlation between ι and P_0: most houses will have a base period price such that $P_{i0} > \tilde{P}_0$, but some houses with $P_{i0} < \tilde{P}_0$ will select in $\mathcal{S}$ because of large positive errors. On the other hand, as P_0 increases, just about any value for ι is conceivable, including very small values. This negative correlation may lead to an underestimated index.

In these circumstances, it is useful to know the determinants of selection in $\mathcal{S}$, as well as their magnitude. This is the purpose of equation (13). There are numerous possible determinants, represented by the vector $\mathbf{n}_i$; I focus on local demographic, economic and housing data, as specified in section (4). Knowing these determinants will enable us to distinguish houses which are and should be in $\mathcal{S}$, based on observable characteristics, from houses which are spuriously in $\mathcal{S}$, *i.e.* houses with large positive values for v_i. Under the hypotheses of the three-equation model, the correlation between ι_i and v_i allows us to control for the selection induced-endogeneity of ι_i. The next subsection details how this can be

[2] Where p_{at_1} is shorthand for $e^{p_{a1}}$, as is ι_i for e^{ι_i}.

done.

3.2 Estimation in the context of sample selection

There are several conceivable options for estimating the parameters of interest, given the model presented in section (3.1). One choice may be to make the additional assumption of trivariate normality of the error terms (ι_i, u_i, v_i), and to estimate the model by maximum likelihood. This is feasible, but computationally demanding, and less robust than a procedure based on Heckman's correction if we allow for correlation between the error terms in the correlation matrix.[3]

The method we used relies on the well known property of bivariate normal variables, allowing us to express ι_i in terms of v_i:

$$\iota_i = \rho\sigma_{\iota_i} v_i + \psi_i$$

where ψ_i is a random normal variable, with mean 0 and variance $1 - \rho^2$. Consequently, using equations (11) and (13), we have

$$\begin{aligned}
\mathbb{E}(y_i|\mathbf{x}_i, \mu_i = 1) &= \mathbf{x}_i\beta + \rho\sigma_{\iota_i}\mathbb{E}(v_i|\mathbf{x}_i, \mu_i = 1) \\
&= \mathbf{x}_i\beta + \rho\sigma_{\iota_i}\mathbb{E}(v_i|\mathbf{x}_i, v_i \geq -\mathbf{n}_i\delta) \\
&= \mathbf{x}_i\beta + \rho\sigma_{\iota_i}\frac{\phi(\mathbf{n}_i\delta)}{\Phi(\mathbf{n}_i\delta)} \qquad (15)
\end{aligned}$$

where the fraction in the last line is the inverse Mills ratio, denoted $\lambda(\mathbf{n}_i\delta)$ for simplicity. It appears from this equation that – since the inverse Mills ratio is strictly positive –, if ρ is positive, the regression line of y_i on $\mathbf{x}_i$ will have a positive bias, and conversely, if ρ is negative, the regression line will have a negative bias. However, in our setup, knowing the sign of ρ does not suffice to predict the estimator's bias for each period. Indeed, ρ interacts with another term whose sign is ambiguous: the product between the matrix $(\mathbf{Z}'\mathbf{X})^{-1}\mathbf{Z}'$ and the vector of inverse Mills ratios. The expression of the non-corrected estimator's expected value, $\mathbb{E}(\hat{\beta}_{\text{ARS}}|\mathbf{X}, \mathbf{Z}, \mathbf{n})$, demonstrates this fairly clearly. Assume there are n sales pairs in $\mathcal{S}$. Let $\boldsymbol{\lambda}$ be a vector of inverse Mills ratios of length n, for the houses in $\mathcal{S}$: entry i is $\lambda(\mathbf{n}_i\delta)$. We then have that

$$\begin{aligned}
\mathbb{E}(\hat{\beta}_{\text{ARS}}|\mathbf{X}, \mathbf{Z}, \mathbf{n}) &= \mathbb{E}((\mathbf{Z}'\mathbf{X})^{-1}\mathbf{Z}'\mathbf{y}|\mathbf{X}, \mathbf{Z}, \mathbf{n}) \\
&= (\mathbf{Z}'\mathbf{X})^{-1}\mathbf{Z}'\mathbf{X}\beta + (\mathbf{Z}'\mathbf{X})^{-1}\mathbf{Z}'\mathbb{E}(\boldsymbol{\iota}|\mathbf{X}, \mathbf{Z}, \mathbf{n}) \\
&= \beta + \rho\sigma(\mathbf{Z}'\mathbf{X})^{-1}\mathbf{Z}'\boldsymbol{\lambda} \qquad (16)
\end{aligned}$$

The second term in equality (16) is a vector of length T. Since the index in period t is the reciprocal of the tth entry of $\hat{\beta}_{\text{ARS}}$, it will be positively biased if the tth

[3] See Wooldridge (2010), p. 813.

entry of $\rho\sigma(\mathbf{Z}'\mathbf{X})^{-1}\mathbf{Z}'\boldsymbol{\lambda}$ is negative, and negatively biased in the other case. Note that if the tth entry of $\rho\sigma(\mathbf{Z}'\mathbf{X})^{-1}\mathbf{Z}'\boldsymbol{\lambda}$ is negative and close in absolute value to the tth entry of β[4], bias of the index will be very large, because of the asymptote at $x = 0$ of the function $x \mapsto 1/x$. The impact of a marginal increase of $\lambda(\mathbf{n}_i\delta)$ on the direction of the non-corrected estimator's bias, however, is ambiguous, and would require a deeper exploration of the sample-corrected model.

Importantly for estimation concerns, since $\mathbf{x}$ is not exogenous, the usual two-step Heckman method would not provide consistent estimates of β.[5] Nevertheless, a similar three-step method should provide consistent estimates:

1. Using all observations ($i \in \mathcal{S} \cup \bar{\mathcal{S}}$), estimate δ with a probit regression of μ on $\mathbf{n}$. With this estimate, compute inverse Mills ratios $\lambda(\mathbf{n}_i\hat{\delta})$.

2. For $i \in \mathcal{S}$, estimate $y_i = \mathbf{x_i}\beta + \lambda(\mathbf{n}_i\hat{\delta}) + \iota_i$ by two stage least squares, using $(\mathbf{z}_i, \lambda(\mathbf{n}_i\hat{\delta}))$ as instruments.

3. Correct for heteroscedasticity by performing a weighted least squares version of the two-stage least squares regression in step 2, as described for the non-corrected repeat sales indices.

This method is recommended by Wooldridge (2010)[6], under the assumptions that (a) $(\mathbf{n}_i, \mu_i)$ is always observed, and $(y_i, \mathbf{x}_i)$ is observed when $\mu_i = 1$, (b) (ι_i, v_i) is independent of $\mathbf{n}_i$, (c) v_i is a standard normal random variable, (d) $\mathbb{E}(\iota_i|v_i)$ is proportional to v_i, and (e) $\mathbb{E}(\mathbf{z}_i u_i) = 0$, and the variables contained in $\mathbf{z}_i$ must not only be exogenous variables from the equation of interest. Assumptions (a), (c) and (d) all directly result from setup of the model. As for assumption (e), the first and second parts are the exclusion restriction and the rank condition, required when using instruments (independently of selection). Lastly, assumption (b) is a typical exogeneity condition in Heckman regressions.

An important issue is that normal standard errors have to be adjusted in this procedure, either via bootstrap, or by correcting the covariance matrix obtained in the last step of the procedure. When performing the regressions, I relied on the computation of the covariance matrix proposed by Heckman, and implemented in the R package `SampleSelection`. Robustness checks should ideally be performed *via* bootstrap.

Finally, it is perfectly conceivable, even in a short period, that a given house is sold more than once. This indeed does occur in the data. Suppose for instance

[4] All the entries of β are positive, since house prices are positive.

[5] This method involves obtaining an estimate $\hat{\delta}$ of δ in a first step probit, and then regressing y on $\mathbf{x}$ and $\lambda(\mathbf{n}_i\hat{\delta})$, by OLS.

[6] Wooldridge, *Econometric Analysis of Cross Section and Panel Data*, section 19.6.2.

that the data contain sales prices for house i in three different dates $\tilde{t}_i > t'_i > t_i$. In this situation, I assume there are two sales pairs, one in periods t_i and t'_i, and a second one in periods t'_i and $\tilde{t}_i$. This induces correlation between the sales pairs, since according to the price forming equation (2), $\text{Cov}(p_{i\tilde{t}_i} - p_{it'_i}, p_{it'_i} - p_{it_i}) = -\sigma_n^2$. Though this correlation could bias the index estimates without adopting a generalized least squares strategy, it seems according to Clapp and Giaccotto (1992) that the bias is small.

4 Data for the Chicago metropolitan area

The data come from two main sources: housing data, containing records of transactions on property deed as well as information regarding all houses in metropolitan Chicago, and census data at the block group level.

A metropolitan statistical area is defined by the United States federal government as one or more adjacent counties with at least one urban core, which includes a population of at least 50,000 (when this population is between 10,000 and 50,000, the region is a called a micropolitan statistical area). An urban core is defined by the Census Bureau as contiguous census geographical units – block groups – having a population density of at least 390/km^2, with surrounding units having a density of at least 190/km^2. The Chicago metropolitan statistical area thus covers 14 counties located in 3 states: Illinois, Wisconsin and Indiana. The total population in 2010 was 9,461,105 according to the Census Bureau.[7]

4.1 Single family housing deeds and housing data

The housing data merge two rich sources: records of transactions, also known as deeds, which are legal documents pertaining to property rights, and county tax data, providing us with a list of all houses in the Chicago metropolitan statistical area, along with their location.

The deed data are collected by CoreLogic, a business, property and consumer information firm. The data base on which I worked contains records for all deeds passed in the Chicago metropolitan statistical area, in the years 2000 to 2014. In the United States, deed registration began around 1640, in the Plymouth and Massachusetts Bay colonies, though it was not an English custom. It serves several purposes, including public information and protection of property rights.[8]

A deed is a legal document, which in real estate law confirms, among other things, the transfer of property rights. Only grant deeds were kept, a type of

[7] https://www2.census.gov/programs-surveys/decennial/tables/cph/cph-t/cph-t-2/cph-t-2.xls

[8] Dukeminier and Krier, *Property*, 2002.

deed generally executed by house sellers, and containing the sale price. Within these grant deeds, I cleaned the data from non arms-length transactions, ensuring that the parties to a transaction are independent and on equal footing, and hence that the property is sold at market price. For the same reason, foreclosed houses were dropped, as well as nominal transactions. Following Standard & Poor's methodology, I dropped deed observations occurring less than six months after the previous deed for the same house, and the observations were restricted to single family housing. This guarantees that the price changes aren't affected by house flipping or fraudulent transactions, and that the index doesn't combine differing trends specific to various types of housing units, respectively. This leaves with us with 460,144 transactions, of which 118,070 are sold twice or more, and altogether 154,928 sales pairs. 386,428 houses were sold at least once during the period of observation.

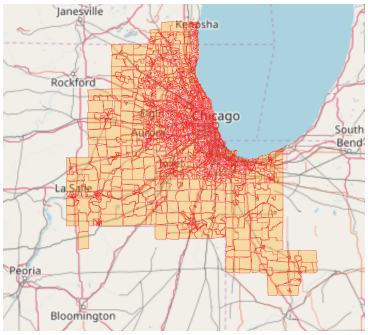
The tax data registers all houses in the Chicago statistical metropolitan area as of 2015, including those which weren't sold during the period of observation. It also contains their geographic coordinates. The data set contains 2,334,649 single family houses. Hence, between 2000 and 2014, 16.6% of all houses were sold at least once.

4.2 Demographic and economic census data

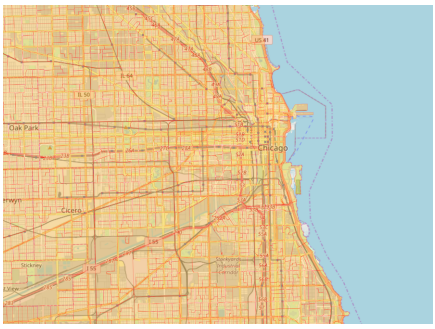
Census data is made publicly available by the National Historical Geographic Information System. I used data at the block group level, which is the smallest geographic unit in Census data; they are constructed to contain a population of 600 to 3,000 people. In the Chicago metropolitan statistical area, there were 6,305 block groups in 2000. So as to merge the housing transaction data with the census data, I made use of the 2000 census block group shape file. Figure (1) displays the Chicago metropolitan statistical area, and a portion of central Chicago. Naturally, since the center denser than the rest of the area, the block groups are smaller.

I focused on variables which are likely to affect the probability of sale frequency, and thus selection in the sample of houses which were sold at least twice during the period of observation, $\mathcal{S}$. Those variables are household median income, the median year in which the block's buildings were built, the fraction of vacant housing, population fractions of different races, population levels, and distance between the block group's centroid and the Central Business District (CBD).

Figure 1: Block group structure in Chicago



(a) Chicago Metropolitan Statistical Area



(b) Central Chicago

It should be noted that Chicago is an intensely segregated city. Though the population is almost evenly distributed among blacks (33%), whites (32%) and Hispanics (29%), most neighborhoods are not highly integrated. For instance, the exposure index of whites to blacks was 4.5% in 2000,[9] meaning that the average fraction of blacks among an average white Chicagoan's neighbors is 4.5%.[10] For individuals who identify as multiracial, a possibility since the 2000 census, segregation appears to be relatively lower: the black/white exposure to whites is 50.7%, and 25.6% to blacks. However, only 0.24% of the Chicago population identifies as black/white. Race is an essential and standard variable in urban economics and urban studies in general, because it captures various underlying local aspects at the neighborhood level. To quote one of many examples, the sociologist Robert J. Sampson makes the case that the percentage of blacks is a strong predictor of the perceived disorder in different Chicago neighborhoods (which is related to mean public behavior that is considered threatening or undesirable, including verbal harassment, garbage in the streets, or violent crime), and is furthermore independent of the respondent's race, as well as a better predictor of perceived disorder than reliable measures of disorder. He goes on to argue that measures of perceived disorder incorporate the stigma of majority black neighborhoods, which may play an important role in the persistence of poverty in many of these neighborhoods.[11] It may thus be expected that perception of neighborhoods based on their racial com-

[9] See Frey and Myers, *Neighborhood Segregation in Single-Race and Multirace America: A Census 2000 Study of Cities and Metropolitan Areas*, 2002.

[10] Let n^b, n_j^b and n_j denote respectively the total number of blacks in the city, the number of blacks in neighborhood j, and the population of neighborhood j; n^w and n_j^w are defined similarly for whites. The exposure index of whites to blacks is then $\sum_j \frac{n_j^w}{n^w}\frac{n_j^w}{n_j}$.

[11] Sampson, *Great American City, Chicago and the Enduring Neighborhood Effect, 2012.*

position impacts the attractiveness of its housing market, not only with respect to prices, but also regarding sale intensity.

The other selection variables are more directly linked to sale intensity, most notably vacant housing and median year of building construction. Indeed, neighborhoods with more vacant housing can be expected to have a larger stock of houses on the housing market, and hence to have more sales; this effect may be ambiguous, however, because very large fractions of vacant housing may be associated with undesirable neighborhood features, and a symptom of its avoidance by buyers. As for the median year of construction, it increases if buildings were newly constructed, and hence on sale. Lastly, distance from the CBD gives information, to some extent, on the benefit of location in a given block group, since most jobs are located in the vicinity of the CBD.

As mentioned earlier, it would be preferable to take into account potential changes across time in the selection variables. This would require a more sophisticated model, and potentially non-existent data: decennial census data is more reliable than the American Community Survey, which is not available before 2005. Hence, local characteristics in the data set at use is set to the year 2000.

4.3 Sale intensity and its relation with selection variables

Before proceeding with the presentation of the results, it is valuable to look at some aspects of the selection variables, and possibly their relationship with housing prices. Even without the repeat sales method, some effects of the 2006 financial crisis should be visible on the prices, as well as on the likelihood to sell. There is indeed substantial heterogeneity between block groups, as made clear in table (3). Variables indicating the presence of the three main ethnic groups (blacks, Hispanics and whites) in neighborhoods, for instance, have large standard deviations, indicating high polarization and segregation levels. The same is true for median income.

Let us focus especially on the ratio of the number of sales to the number of housing units in each block group, for various years. This ratio, referred to as sale intensity, is similar to the sale probability in each block group, though some houses may be sold more than once in a year, virtually making it possible for the ratio to be strictly greater than 1, though this never occurs when it is measured for a single year. Sale intensity is essential, because it may help understand the correction resulting from the estimation strategy presented in subsection (3.2). It is clear from the maps of central Chicago, displayed in figure (2), that there are global and local patterns of sale intensity, though the latter are not as striking. The boom and bust dynamics, preceding and following the 2006 housing crisis, are very apparent.

In the North Side, there is no apparent difference between 2006 and 2014.

Table 3: Summary statistics for the 6,305 block groups

Variable	mean	stand. dev.	median	min	max
Sale intensity (2000-2014)	0.3	0.4	0.1	0	2.1
Median income ($)	53,009.1	26,229.4	48,938	0	200,000
Distance from CBD (miles)	18.4	13.2	14.6	0.1	82.7
Fraction vacant	0.1	0.1	0	0	1
Median year of building const.	1951.3	15	1955	1939	1999
Fraction Asians	0	0.1	0	0	0.9
Fraction blacks	0.2	0.4	0	0	1
Fraction Hispanics	0.2	0.2	0.1	0	1
Fraction whites	0.6	0.4	0.7	0	1
Fraction other races	0	0	0	0	0.3

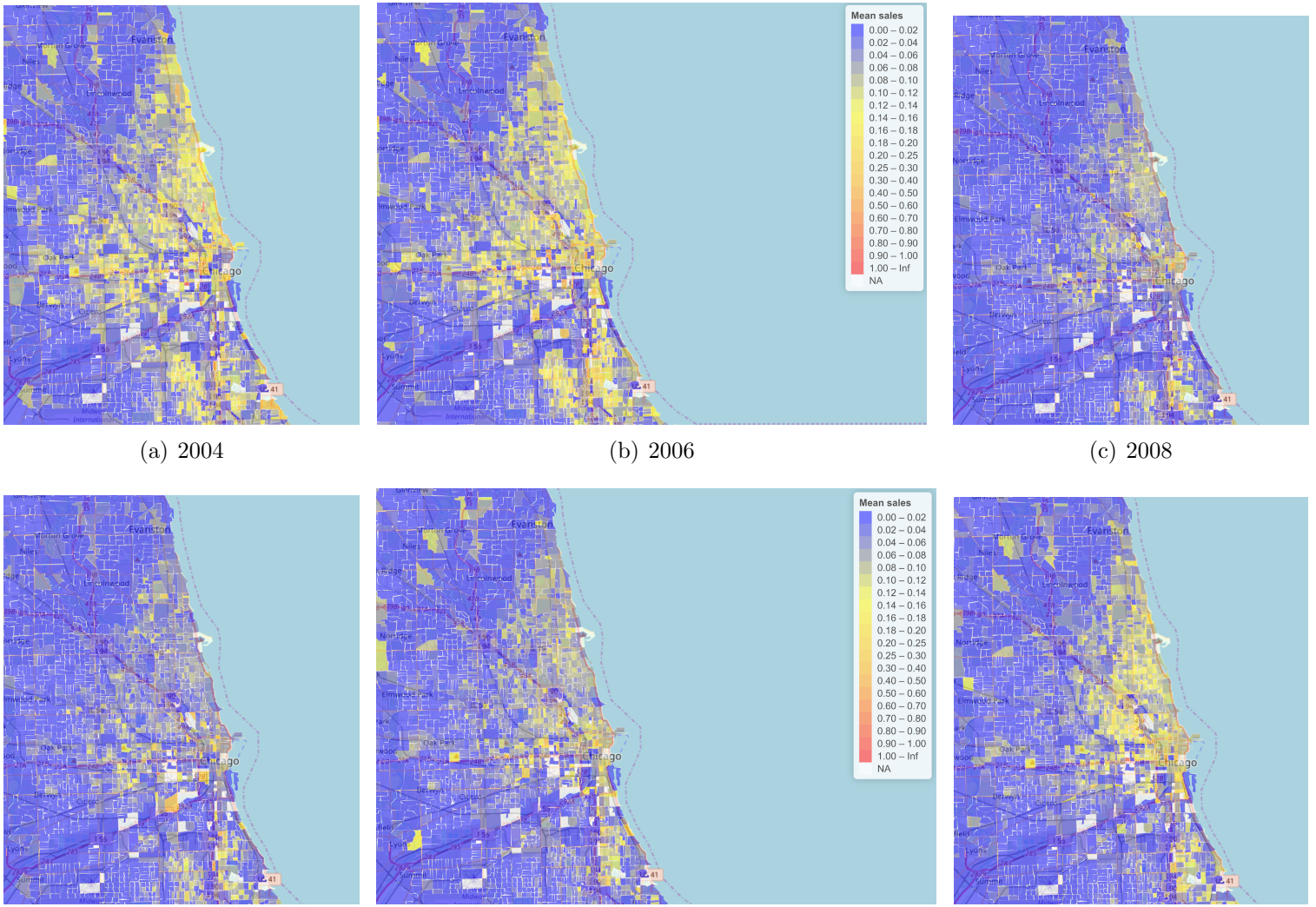
Neighborhoods in the vicinity of the lake and of the CBD maintain a relatively high sale intensity after 2006, whether poor or rich. Particularly rich and white suburbs, such as Wilmette (north of Evanston and thus not visible in the figure), are always characterized by low sale intensity, which don't change much across years.

Unlike the North Side, some neighborhoods of the South Side have strong sales intensities - at least at the beginning of the period - despite not being very close to the center and far from the shore. These neighborhoods are concentrated between West 43d and West 83rd Streets, just west of the University of Chicago. Sales in these neighborhoods were much less dynamic in 2014 than in 2006, though it seems that sale intensity decreased progressively rather than brutally. Though I have not looked for evidence in the data, it is probable that this is due to the many foreclosures that took place in these neighborhoods. Indeed, the southeastern neighborhoods of the South Side culminated in 2007 with a rate above 25 foreclosed houses for every 1,000 mortgageable properties.[12] Foreclosures concentrated in minority areas.

Overall, it seems centrality counteracts bust effects, which lower sale intensity in most places. Central Chicago is not representative of the whole Chicago metropolitan area: in most places, a smaller distance from the center doesn't counteract the negative impacts on sales of high median incomes, or high fractions of whites, for instance.

[12] See Young, *The Foreclosure Crisis in the Chicago Area: Facts, Trends and Responses*, 2008

Figure 2: Sale intensity in block groups from central Chicago



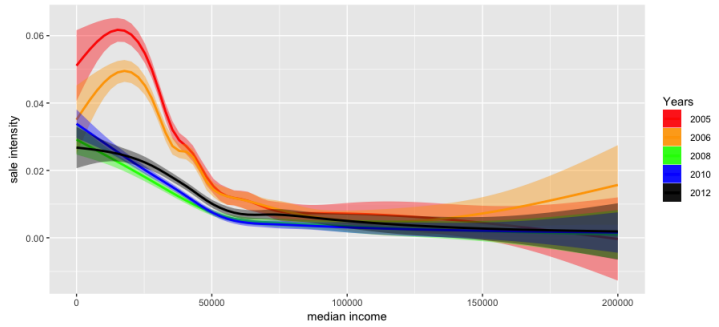
(a) 2004 (b) 2006 (c) 2008

(d) 2010 (e) 2012 (f) 2014

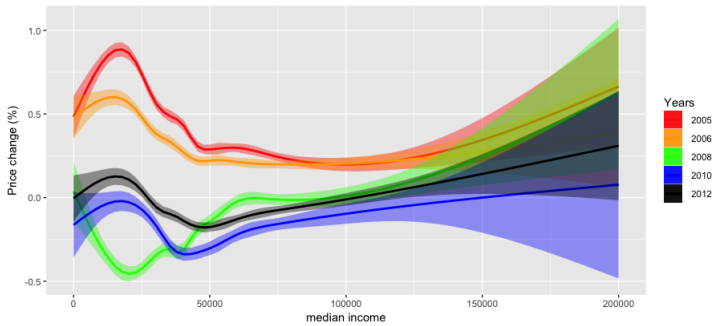
The sale intensity scale is identical in all years.

We now turn to the relationship between sale intensity and the block groups' economic status, as measured by their median income. Panel (a) in figure (3) depicts this relationship for sales occurring between 2005 and 2012. During the whole period, lower income neighborhoods experience higher sale intensity than richer block groups. This is especially true in 2005 and 2006, where there is a peak sale intensity for neighborhoods with a median income just below $25,000. Sale intensity brutally decreases in 2008 – the most severe bust year along with 2007, inducing numerous foreclosures – but still remains higher than richer neighborhoods, and goes on to increase slightly until 2012, in what seems to resemble a recovery period. Panel (b) gives a sense of the percentage changes of nominal prices corresponding to these sales. Quite strikingly, houses in poor neighborhoods had the highest price increase until 2007 (year not displayed in the figure), and subsequently experienced a price decrease in 2008, unlike most houses from neighborhoods with median incomes above $75,000.

Figure 3: Correlations between median income and sales from 2005 to 2012



(a) Sale intensity and median income



(b) Price increase and median income

Graphs produced with the local regression (LOESS) method.

5 Results

The results are presented in two parts, corresponding to the two main stages of the regression. The first part, subsection (5.1), provides a more precise understanding of the determinants of selection in the sample of houses which are sold at least twice. Subsection (5.2) discusses the index itself.

5.1 First step probit

The average marginal effects of the determinants are displayed in the appendix, in table (4), in two different specifications. Because of the apparently non-linear relationship between median income and sale intensity, the impact of median income was estimated on income bins. The same was done for distance from the CBD.

The first specification does not include distance from the CBD. As could be expected, the fraction of vacant houses has a large positive and significant effect on the likelihood of selecting in $\mathcal{S}$, though it falls from 0.42 to 0.11 in the second specification, a sign that omitting centrality causes bias, and that a significant portion of houses in block groups with large fractions of vacant houses were located in central block groups, *i.e.* the ones with the highest sale intensities. Moreover, as was foreshadowed by the graph in panel (a) of figure (3), in the previous subsection, the regression line of selection in $\mathcal{S}$ on median income is concave, reaching a maximum for a median income of about \$20,000. Houses located in neighborhoods with median incomes as high as \$40,000 (in the second specification) still have significantly higher probabilities of being in $\mathcal{S}$ than houses in the poorest block groups (below \$10,000). Starting from a median income of \$60,000, selection less and less likely as the median income increases.

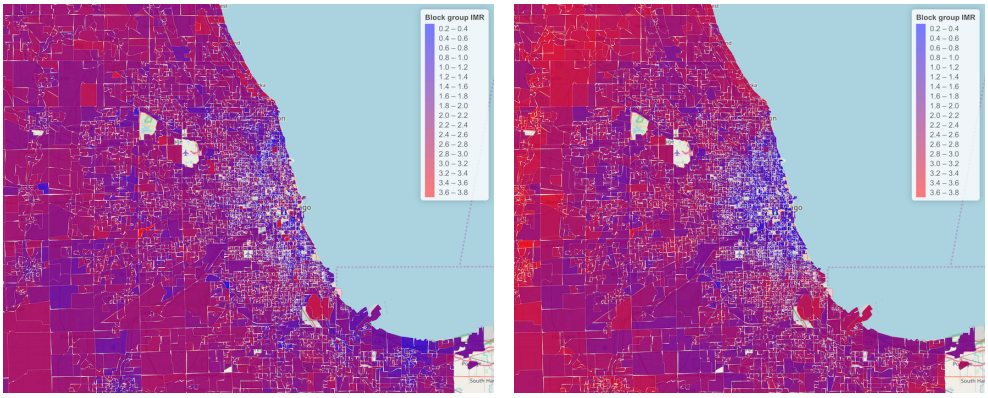
Marginal effects for fractions of ethnic groups are equally informative about selection in $\mathcal{S}$. The fraction of white population was not included in the regressors, to avoid collinearity in the independent variables. The fraction of blacks has a negative effect on selection, and this effect becomes stronger when distance from the CBD is included in the specification, because many block groups with high fractions of blacks are located near the center of Chicago, especially in the South Side. The same is true of Hispanics, whereas selection becomes more likely as the fraction of Asians increases. Surprisingly, selection likelihood increases strongly with the fraction of minorities categorized as "other", but these minorities represent 0% of the population in the average block group.

Selection in $\mathcal{S}$ is, as predicted, less likely as the distance from the CBD increases. This likelihood decreases at an almost constant rate between 4.5 and 10 miles from the CBD, with an average probability decrease of 0.025 for every 2.5 extra mile. The slope of this probability decline is less steep and even positive in

the area between 1 and 4 miles from the CBD, because this area is highly residential, and also more likely to have a higher density of single family homes than the center of Chicago, where large apartment buildings and condos are more common.

It is of central interest to make sense of the inverse Mills ratio (IMR), computed in the first step of the corrected ARS method, just after the probit regression. For instance, are there geographical patterns arising from the IMR values? Given that high IMR values reflect an improbable selection in $\mathcal{S}$ (as long as the only relevant predictors are those contained in $\mathbf{n}$), one should expect that block groups which are farthest from the CBD to have the lowest IMR values. This is indeed the case, as suggested in figure (4). In panel (a), the IMR values correspond to the first regression specification, which did not include distance from CBD. Though it is clear that Chicago and its immediate vicinity have lower IMR values than more distance suburbs, this pattern is more visible in panel (b), where more remote block groups have higher IMR values, indicated by the stronger red hue, and the center is more homogeneously blue. In panel (b), there is a semi-circle, with a radius of about 40 miles and whose center is the CBD, beyond which most block groups have the highest IMR levels (3 and above). How will these changes affect the corrected price index?

Figure 4: IMR values resulting from probit for two different specifications



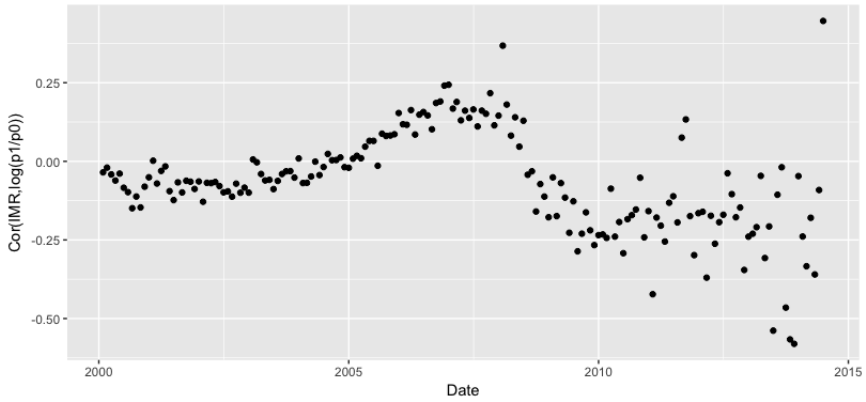
(a) Covariates don't include distance from CBD

(b) Covariates include distance from CBD

A way to address this question is too explore the relationship between IMR levels and price changes. In this respect, one may expect houses which are most likely to select in $\mathcal{S}$ to be characterized by price increases in some periods, but not necessarily in all periods; the same should apply to the houses least likely to select in $\mathcal{S}$. Figure (5) presents the correlation between log price increases and IMR

levels, conditional on time. The IMR values result from the probit specification in which distance from CBD is not included in the covariates. Between 2000 and 2005, and then from the end of 2008 to the end of the period of observation, this correlation is negative. Hence houses which were most likely to select in $\mathcal{S}$ – if the determinants taken into account are restricted to the probit covariates – generally experienced price increases. This was not the case during the most acute boom and bust periods, from 2005 to the end of 2008, when houses which were least likely to select in $\mathcal{S}$ experienced price increases, while the opposite was true for houses which were more typically in $\mathcal{S}$.

Figure 5: Correlation between IMR levels and price change, conditional on time



The date is that of the first sale, for each sale pair. The IMR values were those resulting from specification (1), table (4). When the date is that of the second sale, the graph is similar.

Interestingly, this pattern almost entirely disappears when the IMR values are those from the specification which controls for distance from the CBD: the correlation between IMR values and price increase is negative and almost constant, with an average value of -0.09. The explanation is the following. In the first specification (*i.e.* when distance from the CBD is not in the covariates), one of the unobserved characteristics is distance from the CBD: some block groups with high IMR levels experienced high sale intensity and have numerous houses in $\mathcal{S}$ because of centrality, even though the values taken by the selection variables would predict unlikely selection in $\mathcal{S}$. Such block groups are in central Chicago, and typically have low median incomes. It was shown previously that houses in low median income block groups had high price increases until 2008, at which point they underwent brutal price decreases. The same can be said of neighborhoods with high fractions of blacks and/or Hispanics, but not for block groups with high proportions of whites, which experience smaller price decreases, especially in 2010.[13] Hence it seems that houses from central block groups with high IMR values in the first specification

[13] *Cf* figure (8), in the appendix

are driving the pattern in figure (5). And indeed, when central and low household median income block groups (less than 10 miles from the CBD, and with a median income below $40,000), or central block groups with high fractions of blacks or Hispanics (above 30 percent), are removed from the data, the correlation is mostly negative and almost always below 0.075 from 2005 to the end of 2008. Furthermore, the pattern is maintained when middle to high median incomes ($60,000 and more) are removed from the data.

The correlation between the IMR and price increase entails various potential consequences for the corrected estimator. Consider the case of the first specification. The sign of the estimate of ρ can plausibly be predicted based on this correlation, since during most of the period of observation, high IMR levels are associated with price decrease, and while the opposite is true for roughly a quarter of the total period. Based on the regression equation (15), the relationship between the IMR and price shifts for an individual house can be rewritten

$$\mathbb{E}(P_{it_i}|\mathbf{x}_i, \mu_i)p_{at_i}^{-1} - \mathbb{E}(P_{it'_i}|\mathbf{x}_i, \mu_i)p_{at'_i}^{-1} = \rho\sigma_{\iota_i}\lambda(\mathbf{n}_i\delta)$$

where it is assumed that if $t_i = 0$, then $p_{at_i} = 1$. If price increases are in general negatively related to IMR levels – as suggested thus far –, then ρ should be negative, and conversely if the relation is positive. As explained previously, these two cases are associated with 2^T potential sub-cases, depending on whether the T entries of the vector $(\mathbf{Z}'\mathbf{X})^{-1}\mathbf{Z}'\boldsymbol{\lambda}$ are positive or negative.

As for the second specification, the IMR and price increases are always negatively correlated, with a small absolute value, suggesting that the estimated ρ will be negative and close to 0. Hence the corrected index will most likely be similar to the non-corrected ARS index.

5.2 Corrected *vs* non-corrected index

I present the results from first specification, and subsequently from the second specification, in which the distance from the CBD is included among the covariates. Table (5) in the appendix displays the estimation results.

5.2.1 The corrected index when distance from the CBD is not a selection variable

In this setup, ρ is estimated at -0.0952, and the coefficient for the inverse Mills ratio is -8,394 and highly significant, indicating that if the IMR of house i is 1, then all else being equal, its price for its first sale adjusted for the base period is $8,394 lower than its price in the second sale adjusted for the base period. Indeed, the

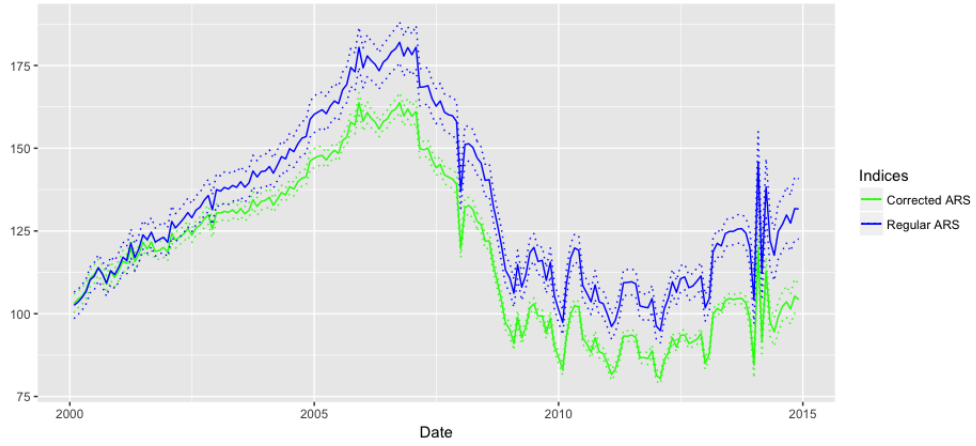
second step regression's structural equation for the corrected ARS can be written

$$P_{it_i}p_{at_i}^{-1} = P_{it'_i}p_{at'_i}^{-1} + \rho\sigma_{\iota_i}\lambda(\mathbf{n}_i\delta) + \iota_i$$

where $\rho\sigma_{\iota_i}$ is the coefficient for the IMR. It follows that houses which are least likely to select in $\mathcal{S}$ – according to the first specification – experienced higher price increases. This confirms evidence from the previous subsection.

The corrected index is displayed in figure (6), which also includes the non-corrected ARS index. Since the 95% confidence intervals do not intersect during most of the period of observation, the difference between the two indices is generally significant. Unsurprisingly, the general trends are comparable for the two indices: prices increase from 2000 to mid-2007, and then decrease until mid-2012, after which they increase again. The shape of the curve reflects the housing boom and bust of this period. Crucially, the corrected index is almost systematically smaller than the regular ARS estimate.

Figure 6: Corrected ARS resulting from the first specification, and regular ARS



The dotted lines are the 95% confidence intervals; the standard errors are adjusted with the delta-method.

A straightforward but incomplete interpretation is that the regular ARS overestimates the boom because of houses with high IMR levels in $\mathcal{S}$, which capture a combination of centrality, low median incomes and a high presence of blacks or Hispanics. Moreover, from 2005 to the end of 2008 – the end of the boom and beginning of the bust periods –, IMR levels were positively correlated with price increases. Hence the housing price boom as depicted by the regular ARS may be partly excessive, since it results from the specific characteristics of houses in block groups with high IMR levels.

However, it was shown that during most of the period of observation, IMR levels and price increases were negatively correlated, so the previous argument does not

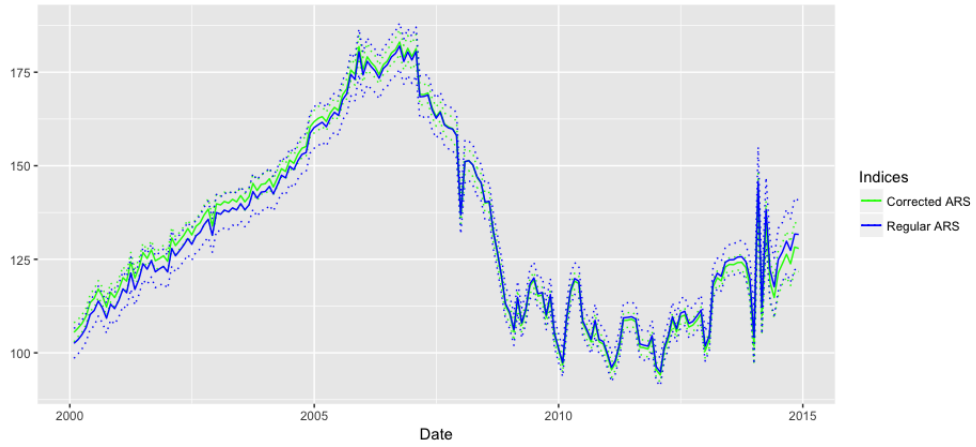
explain why the regular ARS overestimates house values during most of the period. Furthermore, recall that the econometric explanation for the overestimation of the index is that $\rho\sigma(\mathbf{Z}'\mathbf{X})^{-1}\mathbf{Z}'\boldsymbol{\lambda}$ is negative, implying that $(\mathbf{Z}'\mathbf{X})^{-1}\mathbf{Z}'\boldsymbol{\lambda}$ is positive. My current knowledge of the correction model does not enable me to conclude as to what this entails regarding the relationship between $\mathbf{X}$, $\mathbf{Z}$ and $\boldsymbol{\lambda}$.

5.2.2 The corrected index when distance from the CBD is included in the covariates

In this setup, ρ is still negative, though significantly smaller: it is estimated at -0.025. It is thus not surprising that the correlation between the IMR level and the price increase – though still negative and significant – has a smaller magnitude: an IMR of 1 indicates that the base period is \$1,918 lower than its price in the second sale adjusted for the base period.

This low correlation results in a corrected index which is very close to the non-corrected ARS, as displayed in figure (7). The fact that their confidence intervals overlap is not sufficient to conclude that they are not significantly different; however, the confidence intervals adjusted for the difference of the estimates systematically include 0.

Figure 7: Corrected ARS resulting from the second specification, and regular ARS



The dotted lines are the 95% confidence intervals; the standard errors are adjusted with the delta-method.

Hence, when the distance from the CBD is included in the first step selection regression, the corrected ARS index is no longer different from the regular ARS. The econometric reason is that the correlation between the IMR and price changes is very weak. This reflects the fact that houses with high IMR levels are now almost exclusively houses in the outskirts of Chicago, where sale intensity is predominantly low and price changes do not radically differ from the average changes.

6 Conclusion

This study has revealed diverging patterns of housing sale intensity depending on location and demographic variables. It furthermore explored how these differences could be associated with housing price changes, and how they potentially could be integrated within an econometric framework to capture and correct changes in housing values.

It was found that results are very different depending on the inclusion of the distance from the CBD in the covariates of the selection equation: when it is included, the corrected price index does not differ from the regular ARS, whereas the latter is shown to be overestimated when distance from the CBD is not a covariate. A modest conclusion from these diverging results is that housing features and trends within central Chicago are very diverse, but when compared with suburban neighborhoods, their centrality makes them more alike than distinct, because of similar sale intensities. The major conclusion that can be drawn from this difference is that further exploration should make use of more variables characterizing block groups: crime and indicators of various amenities, for instance, may prove useful in exploring the heterogeneity of determinants of selection in the sample.

Finally, a deeper understanding of the correction model may prove useful. First regarding the formula of the corrected index: I took a few steps in this direction, but found so for no obvious clarification as opposed to the normal equations provided by the GRS and ARS indices. Second, the meaning of the bias of the non-corrected index in terms of price changes. And third, an economic model may ideally help shed light on the selection process.

7 References

Bailey , M. J., R. F. Muth, and H. O. Nourse (1963). *A Regression Method for Real Estate Price Index Construction.* Journal of the American Statistical Association.

Clapp, John M., and Carmelo Giacotto (1992) *Estimating Price Trends for Residential Property: A Comparison of Repeat Sales and Assessed Value Methods.* Journal of the American Statistical Association.

Case, Karl E. and Robert J. Shiller (1987). *Prices of Single Family Homes Since 1970: New Indexes for Four Cities.* New England Review of Economics.

Dukemenier, Jesse and James E. Krier (2002). *Property.* Wolters Kluwer Law & Business.

Frey, William H. and Dowell Myers (2002). *Neighborhood Segregation in Single-Race and Multirace America: A Census 2000 Study of Cities and Metropolitan Areas.* Fanny Mae Foundation.

Heckman, James J. (1976). *The Common Structure of Statistical Models of Truncation, Sample Selection and Limited Dependent Variables and a Simple Estimator for Such Models.* Annals of Economic Social Measurement.

Heckman, James J. (1979). *Sample Selection Bias as a Specification Error.* Econometrica.

Sampson, Robert J. (2012). *Great American City. Chicago and the enduring neighborhood effect.* The University of Chicago Press.

Shiller, Robert J. (1990). *Arithmetic Repeat Sales Price Estimators.* Journal of Housing Economics.

Wooldridge, Jeffrey (2010). *Econometric Analysis of Cross Section and Panel Data.* MIT Press.

Young, Stacy (2008). *The Foreclosure Crisis in the Chicago Area: Facts, Trends and Responses.* Working document.

8 Appendix

Table 4: Determinants of selection in $\mathcal{S}$: average marginal effects in the first stage probit

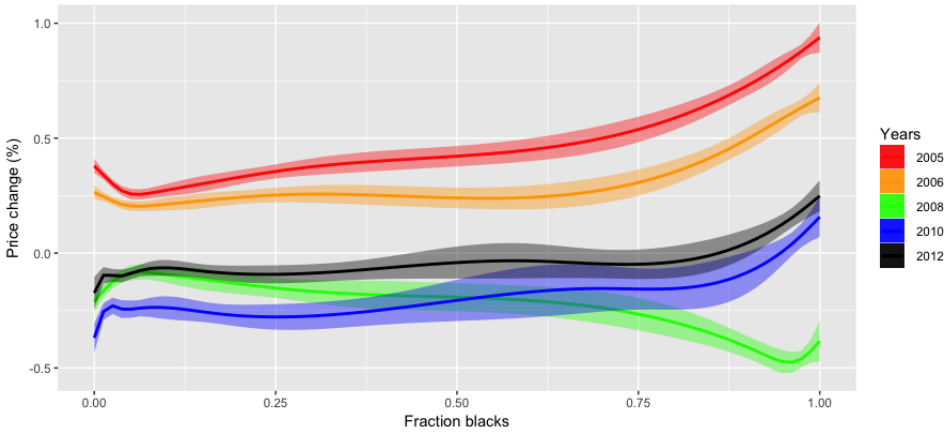
	(1)	(2)
Fraction black	-0.007***	-0.021***
	(0.001)	(-0.001)
Fraction asian	0.23***	0.038***
	(0.002)	(-0.002)
Fraction hispanic	-0.014***	-0.06***
	(0.001)	(-0.001)
Fraction other	1.037***	0.493***
	(0.012)	(-0.011)
Fraction vacant	0.424***	0.115***
	(0.003)	(-0.003)
Median year of building construction	-0.001***	0.001***
	(0)	(0)
Log(population)	0.021***	0.017***
	(0)	(0)
Median income ($)		
(1e+04,2e+04]	0.067***	0.044***
	(0.005)	(-0.002)
(2e+04,3e+04]	0.03***	0.043***
	(0.004)	(-0.002)
(3e+04,4e+04]	0.005*	0.031***
	(0.003)	(-0.002)
(4e+04,5e+04]	0.001	0.025***
	(0.003)	(-0.002)
(5e+04,6e+04]	-0.016***	0.006*
	(0.003)	(-0.002)
(6e+04,1e+05]	-0.019***	-0.013***
	(0.003)	(-0.002)
(1e+05,1.25e+05]	-0.019***	-0.021***
	(0.002)	(-0.002)
(1.25e+05,1.5e+05]	-0.028***	-0.025***
	(0.002)	(-0.003)
(1.5e+05,1.75e+05]	-0.006	-0.026***
	(0.003)	(-0.003)
(1.75e+05,2e+05]	-0.054***	-0.028***
	(0.001)	(-0.003)
Distance from CBD (miles)		
(0.5,1]		-0.187***
		(-0.008)
(1,1.5]		-0.207***
		(-0.008)
(1.5,2]		-0.274***
		(-0.008)
(2,2.5]		-0.202***
		(-0.009)
(2.5,3]		-0.221***
		(-0.008)
(3,3.5]		-0.161***
		(-0.008)
(3.5,4]		-0.186***
		(-0.008)
(4,4.5]		-0.169***

Table 4: Determinants of selection in $\mathcal{S}$: average marginal effects in the first stage probit

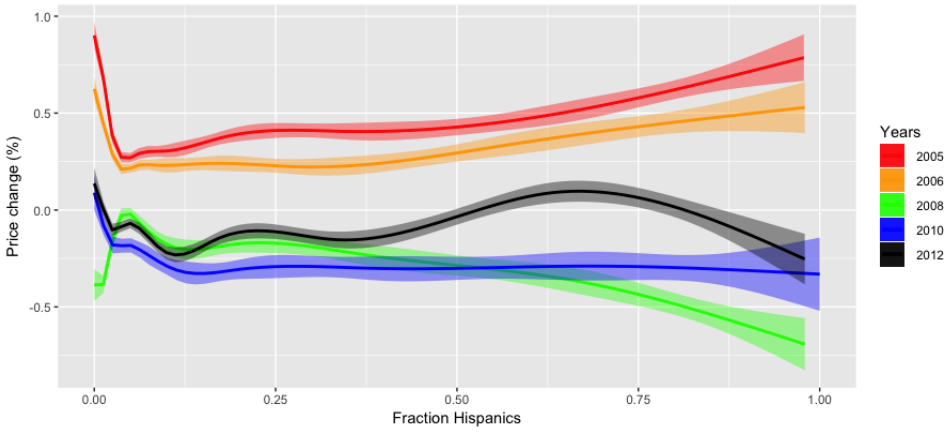
	(1)	(2)
		(-0.008)
(4.5,5]		-0.209***
		(-0.008)
(5,5.5]		-0.206***
		(-0.008)
(5.5,6]		-0.254***
		(-0.008)
(6,6.5]		-0.269***
		(-0.008)
(6.5,7]		-0.283***
		(-0.008)
(7,7.5]		-0.333***
		(-0.008)
(7.5,8]		-0.37***
		(-0.008)
(8,8.5]		-0.387***
		(-0.008)
(8.5,9]		-0.396***
		(-0.008)
(9,9.5]		-0.419***
		(-0.008)
(9.5,10]		-0.44***
		(-0.008)
(10,15]		-0.462***
		(-0.007)
(15,20]		-0.47***
		(-0.007)
(20,25]		-0.466***
		(-0.007)
(25,30]		-0.482***
		(-0.007)
(30,35]		-0.5***
		(-0.007)
(35,40]		-0.502***
		(-0.007)
(40,45]		-0.496***
		(-0.007)
(45,50]		-0.5***
		(-0.007)
(50,Inf]		-0.506***
		(-0.007)

Standard errors are in parentheses. *** and * indicate significance at the 5 percent and 0.1 percent levels, respectively.

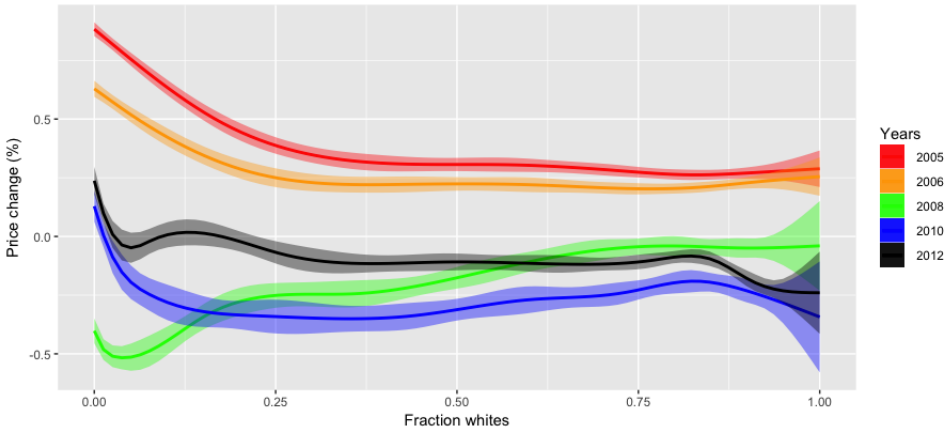
Figure 8: Correlations between fractions of ethnic groups and price changes from 2006 to 2012



(a)



(b)



(c)

Graphs produced with the local regression (LOESS) method.

Table 5: ARS results

		(1)			(2)			Non-corrected ARS		
Year	Month	$\hat{\beta}$	$100\hat{p}_{at}$	std. error	$\hat{\beta}$	$100\hat{p}_{at}$	std. error	$\hat{\beta}_{\text{ARS}}$	std. error	ARS
2000	February	0.9703***	103.06	0.013	0.948***	105.48	0.012	0.975297***	0.0194	102.53
	March	0.9587***	104.31	0.012	0.9384***	106.56	0.010	0.966348***	0.018141	103.48
	April	0.9497***	105.30	0.012	0.9272***	107.85	0.010	0.953461***	0.017814	104.88
	May	0.9339***	107.08	0.011	0.911***	109.77	0.009	0.937182***	0.016984	106.70
	June	0.9036***	110.67	0.010	0.8816***	113.43	0.009	0.907072***	0.016421	110.25
	July	0.8933***	111.94	0.011	0.8729***	114.56	0.010	0.898454***	0.017047	111.30
	August	0.8771***	114.01	0.010	0.8542***	117.07	0.008	0.878405***	0.015744	113.84
	September	0.8921***	112.09	0.010	0.8684***	115.15	0.009	0.893321***	0.0162	111.94
	October	0.9186***	108.86	0.011	0.8909***	112.25	0.010	0.915341***	0.017008	109.25
	November	0.8859***	112.88	0.011	0.8609***	116.16	0.010	0.884919***	0.016896	113.00
	December	0.9013***	110.95	0.011	0.8714***	114.76	0.010	0.894401***	0.016987	111.81
2001	January	0.8843***	113.08	0.012	0.8548***	116.99	0.011	0.876855***	0.017498	114.04
	February	0.8624***	115.95	0.012	0.8328***	120.08	0.011	0.854119***	0.017137	117.08
	March	0.8681***	115.19	0.011	0.8393***	119.15	0.009	0.861586***	0.016069	116.06
	April	0.8357***	119.66	0.010	0.8043***	124.33	0.008	0.824295***	0.01494	121.32
	May	0.8642***	115.71	0.010	0.8333***	120.00	0.008	0.855125***	0.015449	116.94
	June	0.8467***	118.11	0.009	0.814***	122.85	0.008	0.835148***	0.014776	119.74
	July	0.8224***	121.59	0.009	0.7886***	126.81	0.008	0.807667***	0.014476	123.81
	August	0.8365***	119.55	0.008	0.7986***	125.22	0.007	0.817327***	0.013984	122.35
	September	0.8219***	121.67	0.009	0.7841***	127.53	0.008	0.802218***	0.014388	124.65
	October	0.8417***	118.81	0.009	0.8024***	124.63	0.008	0.821998***	0.014626	121.66
	November	0.8395***	119.12	0.009	0.7987***	125.20	0.008	0.816639***	0.014782	122.45
	December	0.8331***	120.03	0.011	0.794***	125.95	0.009	0.812399***	0.015538	123.09
2002	January	0.8427***	118.67	0.010	0.8035***	124.46	0.009	0.822279***	0.015108	121.61
	February	0.8044***	124.32	0.010	0.7651***	130.70	0.009	0.78198***	0.014753	127.88
	March	0.8181***	122.23	0.008	0.7772***	128.67	0.007	0.794086***	0.013838	125.93
	April	0.8119***	123.17	0.008	0.7691***	130.02	0.007	0.785281***	0.013361	127.34
	May	0.8059***	124.08	0.008	0.7611***	131.39	0.007	0.776577***	0.013239	128.77
	June	0.7957***	125.68	0.008	0.7507***	133.21	0.007	0.766022***	0.013061	130.54
	July	0.8085***	123.69	0.007	0.7604***	131.51	0.006	0.775175***	0.013021	129.00
	August	0.7942***	125.91	0.007	0.7473***	133.81	0.006	0.761748***	0.012845	131.28

Table 5: ARS results

		(1)			(2)			Non-corrected ARS		
Year	Month	$\hat{\beta}$	$100\hat{p}_{at}$	std. error	$\hat{\beta}$	$100\hat{p}_{at}$	std. error	$\hat{\beta}_{\text{ARS}}$	std. error	ARS
	September	0.7889***	126.76	0.008	0.7421***	134.75	0.007	0.756391***	0.013073	132.21
	October	0.7777***	128.58	0.008	0.7307***	136.85	0.007	0.744792***	0.013189	134.27
	November	0.7699***	129.89	0.009	0.7225***	138.41	0.007	0.736846***	0.013453	135.71
	December	0.7966***	125.53	0.008	0.747***	133.87	0.007	0.760959***	0.013341	131.41
2003	January	0.7667***	130.43	0.008	0.7152***	139.82	0.007	0.727318***	0.013084	137.49
	February	0.7666***	130.45	0.007	0.7166***	139.55	0.006	0.729743***	0.01245	137.03
	March	0.7638***	130.92	0.007	0.712***	140.45	0.006	0.723815***	0.012466	138.16
	April	0.7656***	130.62	0.007	0.7141***	140.04	0.006	0.726343***	0.012536	137.68
	May	0.7612***	131.37	0.007	0.7091***	141.02	0.006	0.720755***	0.012208	138.74
	June	0.7673***	130.33	0.007	0.7125***	140.35	0.006	0.723697***	0.012036	138.18
	July	0.7589***	131.77	0.007	0.704***	142.05	0.006	0.714909***	0.011983	139.88
	August	0.7681***	130.19	0.007	0.7125***	140.35	0.006	0.723523***	0.011992	138.21
	September	0.7621***	131.22	0.006	0.7057***	141.70	0.006	0.716813***	0.011778	139.51
	October	0.7438***	134.44	0.007	0.6883***	145.28	0.006	0.698758***	0.01189	143.11
	November	0.7553***	132.40	0.008	0.6974***	143.39	0.007	0.707536***	0.012468	141.34
	December	0.7467***	133.92	0.007	0.6896***	145.01	0.006	0.69959***	0.012089	142.94
2004	January	0.7449***	134.25	0.008	0.6886***	145.22	0.007	0.698687***	0.012428	143.13
	February	0.7398***	135.17	0.008	0.6825***	146.52	0.007	0.692178***	0.012323	144.47
	March	0.7532***	132.77	0.006	0.6929***	144.32	0.005	0.701943***	0.011498	142.46
	April	0.7429***	134.61	0.006	0.6813***	146.78	0.005	0.690134***	0.01125	144.90
	May	0.732***	136.61	0.006	0.6702***	149.21	0.005	0.678169***	0.011126	147.46
	June	0.7359***	135.89	0.005	0.6735***	148.48	0.005	0.681341***	0.010896	146.77
	July	0.7227***	138.37	0.006	0.66***	151.52	0.005	0.667331***	0.01077	149.85
	August	0.7264***	137.67	0.006	0.6638***	150.65	0.005	0.671368***	0.01089	148.95
	September	0.715***	139.86	0.006	0.6532***	153.09	0.005	0.660716***	0.010865	151.35
	October	0.7095***	140.94	0.006	0.6465***	154.68	0.005	0.653453***	0.010956	153.03
	November	0.7075***	141.34	0.006	0.6445***	155.16	0.005	0.651325***	0.010942	153.53
	December	0.6843***	146.13	0.006	0.6233***	160.44	0.006	0.629984***	0.010851	158.73
2005	January	0.6803***	146.99	0.007	0.618***	161.81	0.006	0.624041***	0.010999	160.25
	February	0.6774***	147.62	0.007	0.6147***	162.68	0.006	0.621174***	0.011106	160.99
	March	0.6766***	147.80	0.006	0.613***	163.13	0.005	0.61873***	0.010347	161.62

Table 5: ARS results

		(1)			(2)			Non-corrected ARS		
Year	Month	$\hat{\beta}$	$100\hat{p}_{at}$	std. error	$\hat{\beta}$	$100\hat{p}_{at}$	std. error	$\hat{\beta}_{\text{ARS}}$	std. error	ARS
	April	0.683***	146.41	0.006	0.6179***	161.84	0.005	0.623347***	0.010223	160.42
	May	0.6748***	148.19	0.005	0.6092***	164.15	0.005	0.614417***	0.009987	162.76
	June	0.6695***	149.37	0.005	0.6039***	165.59	0.004	0.608728***	0.009809	164.28
	July	0.6737***	148.43	0.005	0.607***	164.75	0.005	0.611789***	0.009948	163.46
	August	0.6565***	152.32	0.005	0.5918***	168.98	0.004	0.596478***	0.009688	167.65
	September	0.6521***	153.35	0.005	0.5863***	170.56	0.005	0.590843***	0.009753	169.25
	October	0.6333***	157.90	0.005	0.5695***	175.59	0.005	0.573347***	0.00962	174.41
	November	0.6371***	156.96	0.006	0.5733***	174.43	0.005	0.577801***	0.009703	173.07
	December	0.6107***	163.75	0.006	0.5499***	181.85	0.005	0.554089***	0.009525	180.48
2006	January	0.6322***	158.18	0.006	0.5694***	175.62	0.005	0.573849***	0.010091	174.26
	February	0.6219***	160.80	0.007	0.5583***	179.12	0.006	0.562111***	0.010146	177.90
	March	0.6286***	159.08	0.006	0.563***	177.62	0.005	0.566482***	0.009658	176.53
	April	0.6333***	157.90	0.006	0.5667***	176.46	0.005	0.570375***	0.009574	175.32
	May	0.6419***	155.79	0.005	0.5734***	174.40	0.004	0.576746***	0.009306	173.39
	June	0.6327***	158.05	0.005	0.565***	176.99	0.004	0.568219***	0.009197	175.99
	July	0.6296***	158.83	0.005	0.5618***	178.00	0.005	0.564858***	0.009412	177.04
	August	0.6208***	161.08	0.005	0.5551***	180.15	0.004	0.558385***	0.009153	179.09
	September	0.6181***	161.79	0.006	0.5523***	181.06	0.005	0.555405***	0.00947	180.05
	October	0.6106***	163.77	0.006	0.5462***	183.08	0.005	0.549429***	0.009477	182.01
	November	0.626***	159.74	0.006	0.5591***	178.86	0.005	0.562266***	0.009759	177.85
	December	0.6177***	161.89	0.006	0.5513***	181.39	0.005	0.554283***	0.009787	180.41
2007	January	0.6265***	159.62	0.006	0.5581***	179.18	0.006	0.560982***	0.010027	178.26
	February	0.621***	161.03	0.007	0.5516***	181.29	0.006	0.553996***	0.010224	180.51
	March	0.6674***	149.83	0.006	0.5914***	169.09	0.005	0.593766***	0.010101	168.42
	April	0.6687***	149.54	0.005	0.5917***	169.00	0.005	0.593351***	0.009888	168.53
	May	0.6665***	150.04	0.005	0.5902***	169.43	0.005	0.592036***	0.00985	168.91
	June	0.6825***	146.52	0.005	0.6042***	165.51	0.005	0.606342***	0.009918	164.92
	July	0.6941***	144.07	0.006	0.6129***	163.16	0.005	0.614651***	0.010222	162.69
	August	0.6893***	145.07	0.006	0.6076***	164.58	0.005	0.608726***	0.010135	164.28
	September	0.7045***	141.94	0.007	0.6205***	161.16	0.006	0.621492***	0.01088	160.90
	October	0.7077***	141.30	0.007	0.6235***	160.38	0.006	0.62457***	0.01114	160.11

Table 5: ARS results

		(1)			(2)			Non-corrected ARS		
Year	Month	$\hat{\beta}$	$100\hat{p}_{at}$	std. error	$\hat{\beta}$	$100\hat{p}_{at}$	std. error	$\hat{\beta}_{\text{ARS}}$	std. error	ARS
	November	0.7109***	140.67	0.007	0.6252***	159.95	0.006	0.625802***	0.011402	159.79
	December	0.7189***	139.10	0.009	0.6322***	158.18	0.007	0.632738***	0.012237	158.04
2008	January	0.8363***	119.57	0.009	0.7314***	136.72	0.008	0.73055***	0.013867	136.88
	February	0.7574***	132.03	0.009	0.6622***	151.01	0.008	0.661642***	0.012804	151.14
	March	0.7534***	132.73	0.007	0.6604***	151.42	0.006	0.66074***	0.011561	151.34
	April	0.7606***	131.47	0.008	0.666***	150.15	0.007	0.665663***	0.012298	150.23
	May	0.7811***	128.03	0.007	0.6809***	146.86	0.006	0.679619***	0.011869	147.14
	June	0.7878***	126.94	0.007	0.6879***	145.37	0.006	0.687295***	0.011744	145.50
	July	0.8204***	121.89	0.008	0.7142***	140.02	0.007	0.712777***	0.012676	140.30
	August	0.8194***	122.04	0.008	0.7131***	140.23	0.007	0.711694***	0.01278	140.51
	September	0.8731***	114.53	0.009	0.7579***	131.94	0.008	0.755405***	0.013876	132.38
	October	0.9136***	109.46	0.010	0.7908***	126.45	0.008	0.78795***	0.014768	126.91
	November	0.9641***	103.72	0.012	0.8337***	119.95	0.011	0.830069***	0.017412	120.47
	December	1.029***	97.18	0.011	0.8887***	112.52	0.010	0.884161***	0.017476	113.10
2009	January	1.049***	95.33	0.012	0.9061***	110.36	0.010	0.901983***	0.018271	110.87
	February	1.099***	90.99	0.012	0.947***	105.60	0.011	0.941724***	0.019042	106.19
	March	1.01***	99.01	0.010	0.8744***	114.36	0.009	0.870865***	0.016741	114.83
	April	1.079***	92.68	0.011	0.9307***	107.45	0.009	0.925835***	0.017564	108.01
	May	1.044***	95.78	0.010	0.8998***	111.14	0.008	0.895105***	0.016212	111.72
	June	0.9825***	101.78	0.009	0.8484***	117.87	0.007	0.844169***	0.014943	118.46
	July	0.971***	102.99	0.009	0.8371***	119.46	0.008	0.833142***	0.014844	120.03
	August	1.008***	99.21	0.009	0.8684***	115.15	0.008	0.863703***	0.015591	115.78
	September	1.006***	99.40	0.009	0.866***	115.47	0.008	0.861096***	0.015585	116.13
	October	1.062***	94.16	0.009	0.9136***	109.46	0.008	0.908516***	0.016331	110.07
	November	1.013***	98.72	0.010	0.8705***	114.88	0.009	0.865146***	0.016274	115.59
	December	1.108***	90.25	0.010	0.9546***	104.76	0.009	0.949166***	0.017785	105.36
2010	January	1.152***	86.81	0.012	0.9936***	100.64	0.010	0.987944***	0.019321	101.22
	February	1.202***	83.20	0.012	1.034***	96.71	0.011	1.027007***	0.020029	97.37
	March	1.073***	93.20	0.010	0.9229***	108.35	0.009	0.917267***	0.017027	109.02
	April	1.001***	99.90	0.010	0.8612***	116.12	0.008	0.856078***	0.015791	116.81
	May	0.9758***	102.48	0.009	0.8391***	119.17	0.008	0.834304***	0.015049	119.86

Table 5: ARS results

		(1)			(2)			Non-corrected ARS		
Year	Month	$\hat{\beta}$	$100\hat{p}_{at}$	std. error	$\hat{\beta}$	$100\hat{p}_{at}$	std. error	$\hat{\beta}_{\text{ARS}}$	std. error	ARS
	June	0.9806***	101.98	0.009	0.8438***	118.51	0.007	0.839163***	0.014904	119.17
	July	1.08***	92.59	0.011	0.927***	107.88	0.009	0.921213***	0.017737	108.55
	August	1.1***	90.91	0.011	0.9471***	105.58	0.009	0.941897***	0.0178	106.17
	September	1.132***	88.34	0.011	0.9716***	102.92	0.009	0.965314***	0.018259	103.59
	October	1.078***	92.76	0.012	0.9261***	107.98	0.011	0.919929***	0.018657	108.70
	November	1.129***	88.57	0.012	0.9714***	102.94	0.011	0.965466***	0.019441	103.58
	December	1.137***	87.95	0.011	0.9769***	102.36	0.010	0.970635***	0.018556	103.03
2011	January	1.176***	85.03	0.012	1.009***	99.11	0.011	1.001612***	0.020204	99.84
	February	1.225***	81.63	0.013	1.048***	95.42	0.012	1.040308***	0.021303	96.12
	March	1.202***	83.20	0.011	1.029***	97.18	0.010	1.021011***	0.019211	97.94
	April	1.146***	87.26	0.011	0.9836***	101.67	0.010	0.976919***	0.018606	102.36
	May	1.072***	93.28	0.010	0.9203***	108.66	0.008	0.914571***	0.01659	109.34
	June	1.073***	93.20	0.010	0.9197***	108.73	0.008	0.913492***	0.01662	109.47
	July	1.069***	93.55	0.010	0.9175***	108.99	0.009	0.91198***	0.016786	109.65
	August	1.076***	92.94	0.009	0.9232***	108.32	0.008	0.917031***	0.016264	109.05
	September	1.153***	86.73	0.011	0.985***	101.52	0.009	0.976879***	0.018484	102.37
	October	1.153***	86.73	0.012	0.9866***	101.36	0.011	0.979996***	0.019531	102.04
	November	1.157***	86.43	0.012	0.9899***	101.02	0.011	0.982411***	0.019667	101.79
	December	1.125***	88.89	0.011	0.9633***	103.81	0.010	0.956045***	0.018485	104.60
2012	January	1.228***	81.43	0.014	1.047***	95.51	0.012	1.038532***	0.021421	96.29
	February	1.244***	80.39	0.012	1.063***	94.07	0.011	1.054263***	0.020641	94.85
	March	1.164***	85.91	0.010	0.9933***	100.67	0.009	0.985554***	0.017994	101.47
	April	1.134***	88.18	0.010	0.9639***	103.75	0.009	0.955133***	0.017371	104.70
	May	1.079***	92.68	0.009	0.9191***	108.80	0.008	0.912024***	0.015955	109.65
	June	1.112***	89.93	0.009	0.9472***	105.57	0.007	0.939415***	0.016199	106.45
	July	1.07***	93.46	0.009	0.9121***	109.64	0.008	0.90427***	0.016353	110.59
	August	1.068***	93.63	0.009	0.9084***	110.08	0.008	0.89995***	0.015914	111.12
	September	1.099***	90.99	0.010	0.9356***	106.88	0.008	0.927241***	0.016808	107.85
	October	1.093***	91.49	0.009	0.931***	107.41	0.008	0.9226***	0.016402	108.39
	November	1.083***	92.34	0.010	0.9193***	108.78	0.008	0.910054***	0.016791	109.88
	December	1.064***	93.98	0.009	0.9058***	110.40	0.008	0.897436***	0.016304	111.43

Table 5: ARS results

		(1)			(2)			Non-corrected ARS		
Year	Month	$\hat{\beta}$	$100\hat{p}_{at}$	std. error	$\hat{\beta}$	$100\hat{p}_{at}$	std. error	$\hat{\beta}_{\text{ARS}}$	std. error	ARS
2013	January	1.177***	84.96	0.012	0.9947***	100.53	0.010	0.982539***	0.019662	101.78
	February	1.138***	87.87	0.011	0.9659***	103.53	0.010	0.955816***	0.018741	104.62
	March	1.002***	99.80	0.009	0.8502***	117.62	0.007	0.841817***	0.015165	118.79
	April	0.9842***	101.60	0.007	0.833***	120.05	0.006	0.824584***	0.01412	121.27
	May	0.9913***	100.88	0.007	0.8386***	119.25	0.006	0.830076***	0.013975	120.47
	June	0.9614***	104.02	0.007	0.8133***	122.96	0.006	0.805065***	0.013405	124.21
	July	0.9558***	104.62	0.007	0.8089***	123.62	0.006	0.800971***	0.013295	124.85
	August	0.9589***	104.29	0.007	0.8094***	123.55	0.006	0.800659***	0.013443	124.90
	September	0.9564***	104.56	0.008	0.8056***	124.13	0.007	0.796239***	0.014006	125.59
	October	0.9557***	104.63	0.008	0.8052***	124.19	0.007	0.795554***	0.013909	125.70
	November	0.9685***	103.25	0.009	0.8143***	122.80	0.008	0.803984***	0.015127	124.38
	December	1.018***	98.23	0.013	0.8492***	117.76	0.012	0.834528***	0.018828	119.83
2014	January	1.184***	84.46	0.027	0.9826***	101.77	0.024	0.959822***	0.033856	104.19
	February	0.8342***	119.88	0.014	0.6996***	142.94	0.013	0.685599***	0.021881	145.86
	March	1.089***	91.83	0.024	0.9097***	109.93	0.021	0.891549***	0.029007	112.16
	April	0.8855***	112.93	0.016	0.7383***	135.45	0.014	0.723012***	0.02277	138.31
	May	1.025***	97.56	0.023	0.8412***	118.88	0.020	0.818726***	0.027527	122.14
	June	1.058***	94.52	0.024	0.8714***	114.76	0.021	0.849987***	0.028257	117.65
	July	1.01***	99.01	0.024	0.824***	121.36	0.021	0.800563***	0.027459	124.91
	August	0.9789***	102.16	0.021	0.8077***	123.81	0.018	0.787725***	0.024905	126.95
	September	0.9653***	103.59	0.023	0.7915***	126.34	0.020	0.770241***	0.025793	129.83
	October	0.9859***	101.43	0.022	0.8078***	123.79	0.019	0.785147***	0.02653	127.36
	November	0.9495***	105.32	0.024	0.7796***	128.27	0.021	0.758945***	0.028133	131.76
	December	0.9591***	104.26	0.022	0.7818***	127.91	0.019	0.759599***	0.026701	131.65
Inverse Mills Ratio		-8394***		169.400	-1918***		149.700			
	σ_ι	88171.136			76160.000					
	ρ	-0.095			-0.025					
Multiple R-Squared		-28.76			-28.760					
Adjusted R-Squared		-28.7948			-23.773					