The association between public transportation infrastructure and home price growth and stability.

Gregory S. Macfarlane[a,*], Juan Moreno-Cruz[b]

[a] *School of Civil and Environmental Engineering, Georgia Institute of Technology, 790 Atlantic Drive, Atlanta GA 30332-0355*
[b] *School of Economics, Georgia Institute of Technology, 221 Bobby Dodd Way, Atlanta, GA 30332*

Abstract

Public rail transit infrastructure and other features of the urban environment shape housing markets, as neighborhoods with high accessibility also tend to be highly valued. But housing markets are dynamic and sometimes turbulent, and the role that transit infrastructure plays in long-term home value in the face of macroscopic events has not been studied. In this paper, we model the performance of the Atlanta housing market in the period from 2002-2012 as a function of a home's proximity to the MARTA rail network. Univariate spatial Durbin models show homes in proximity to public transit had higher values and higher growth rates over the period than homes further away. Multivariate latent class mixture models confirm these results and also show that the Atlanta housing market can be considered as two distinct classes: homes near to MARTA are more likely to be in a class with positive value growth over the period.

Keywords: spatial econometrics, home price volatility, transportation accessibility

1. Background

The story of the US housing market in the period from 2000 through 2012 can be told in three parts, illustrated by the plot of the Case-Shiller home price index in Figure 1 (Standard & Poors, 2013). From 2000 through 2005, home prices rose sharply, with the average home selling in late 2005 for almost twice its 2000 value. After leveling off through 2006, home prices fell precipitously through 2007 and 2008. The market has remained depressed in the intervening years, though a turning point may have been reached in the winter of 2011-2012. But the story told by the composite Case-Shiller index is not homogeneous across the US. Some markets outperformed the national average, gaining more in value during the boom and losing less during the bust. Others underperformed. In the Atlanta market specifically (also shown in Figure 1), the 2007 apex was substantially lower than the national average, and the nadir was more severe.

*Corresponding author. Tel.: +1 801 616 9822

Email addresses: gregmacfarlane@gatech.edu (Gregory S. Macfarlane), juan.moreno-cruz@econ.gatech.edu (Juan Moreno-Cruz)

The heterogeneity in the response goes deeper. Homes in some neighborhoods remained stable in value throughout the period, others fluctuated wildly, and some others may even have gained value in the first five years of the new century without subsequently losing it.

```
# Case shiller indices
cs <- rbind(
  # atlanta
  read_csv("../data/caseshiller/ATXRSA.csv",
           col_names = c("date", "value"), skip = 1) %>%
  mutate(series = "Atlanta"),
  # usa
  read_csv("../data/caseshiller/USCSCOMHPISA.csv",
           col_names = c("date", "value"), skip = 1) %>%
  mutate(series = "US 20 market composite")
) %>%
  # common dates
  filter(date >= "2000-01-01") %>%
  filter(date <= "2013-01-01")


# Dates of US recessions
recesssions <- data_frame(
  start = as.Date(c("2001-03-01", "2007-12-01")),
  end   = as.Date(c("2001-11-30", "2009-06-30")),
  bottom = -Inf, top = Inf
)
```

[Figure 1 about here.]

This heterogeneity in price performance is not necessarily surprising. Neighborhoods are themselves heterogeneous; every neighborhood has its own unique blend of amenities. Homes after all are consumable goods as much as they are financial investments, and households will pay for the amenities that they value; it is possible that the relative value of these amenities changes with (and perhaps because of) the overall economy. But understanding which neighborhood characteristics correlate with advantageous price performance — and which of these characteristics planners can influence — is an important consideration for urban policy.

One possible amenity is transportation accessibility, and access to public rail transit in particular. Neighborhoods with access to public transportation have a higher value, as identified by both theory Alonso (1960) and empirical observation Lewis-Workman and Brod (1997); Iacono and Levinson (2011). There are also several mechanisms by which public transit accessibility may be associated with better long-term value performance. It may be that transit-supportive development (such as high population densities) constrains the housing supply, preventing overbuilding. It may also be that homes with better access to a region's opportunities are able to adapt more quickly to economic changes; laid-off workers in such homes may be able to find a new job more quickly and avoid relocation or default (as was observed by Pivo (2013)).

This paper presents an empirical investigation of home price growth and volatility in Fulton County, Georgia between 2002 and 2012, testing the theory that public rail transit infrastructure — specifically, the Metropolitan Atlanta Rapid Transit Authority (MARTA) rail network — is correlated with positive growth or stability in a housing market. Spatial Durbin models of average value, average growth, and variance in growth show that neighborhoods closer to MARTA stations have higher average values, higher average growth rates, but higher growth volatility than homes further away. These findings are based on terms interacted with the growth in neighborhood income. Further, latent class mixture models reveal at least two distinct home markets in the Atlanta area; homes nearer to the MARTA network are more likely to belong to a class of homes that experienced positive price growth over the period in question.

The paper proceeds as follows. The rest of this section discusses the relevant literature and places the study a theoretical context. Section 2 describes the dataset used in this analysis, which is constructed from the Fulton County tax assessor's database and other publicly-available sources. Section 3 presents univariate spatial regressions of value, growth, and volatility on transit proximity. Section 4 presents a multivariate latent class membership model to investigate the relationship between several simultaneous home price variable, and considers the hypothesis that more than one housing market exists, segregated by transit access. The study concludes with an interpretation of the findings and an outline for further investigation.

1.1. Literature

Numerous authors have identified a correlation between investment in public transit and increased home values. These studies can generally be classified into two broad types. The first type of study considers home prices in a particular cross-section and measures the market's willingness-to-pay for transit proximity (e.g., Lewis-Workman and Brod, 1997; Bowes and Ihlanfeldt, 2001; Debrezion et al., 2007); these studies of necessity assume that the housing market is in some sort of equilibrium. The second study type measures price changes in response to transit construction, assuming that unobserved or endogenous market characteristics can be differenced out (e.g., Grass, 1992; McMillen and McDonald, 2004). Both types of study have generally shown homes near transit stations are more valuable in equilibrium and also that expanding transit infrastructure results in increased home values. The consequences of these findings are twofold: transit accessibility is valued by the market and therefore it may be possible for governments to recoup the cost of construction and operation through elevated property taxes (Smith and Gihring, 2006); however, elevated property values and concurrent gentrification may displace the very populations who most rely on public transit (Pollack et al., 2010).

In the first type of study the housing market and transit infrastructure are both fixed. In the second type, the housing market is considered fixed[1] as transit infrastructure changes. The existing literature — to the best of our knowledge — is missing a potential third type of study, where the transit network remains constant *as the housing market changes*. In particular, we ask, "are homes in proximity to transit networks *more resilient* to demand-related shocks to the housing market?"

There are economic theories that explain why differentials between and within housing markets exist, and how public transportation infrastructure or other features of the built environment may contribute to this heterogeneity. Glaeser et al. (2008) present a model that predicts inelastic housing markets — those with constraints on construction — will perform better during an exogenous and irrational home price bubble. Specifically, inelastic markets cannot build new homes to meet an imagined demand, and so supply is not artificially inflated. Prices rise in response to demand, but return to their proper value when demand subsides. In elastic markets by contrast, the home supply expands and keeps prices from rising but results in an oversupply and a consequent price collapse when the perceived demand returns to its real level. Home prices in Atlanta rose more modestly than the national average from 2000-2006, and dropped below their 2000 level by mid-2011. This theory is supported by the fact that Atlanta is among the most sprawling cities in the U.S. (Ewing et al., 2014), but there may be multiple submarkets within Atlanta. Considering Fulton County in particular, we might hypothetically identify at least two distinct housing markets: an elastic market in the suburban north and southwest, and an inelastic market in the inner city. In this case, proximity to MARTA rail may serve as an indicator of an urban environment and presumably constrained development. Homes in neighborhoods close to MARTA would show higher values but potentially *more* volatility under this model.

Guerrieri et al. (2013) present another explanatory model of disparate market reaction to changes in demand. This spatial model considers the location of wealthy and poor residents in response to an exogenous increase in demand (for example, a housing boom). The model predicts that wealthy residents displace poor residents in neighborhoods adjacent to existing wealthy neighborhoods. Home values in these gentrifying neighborhoods would therefore increase in value more quickly than the rest of the market during the demand shock. In the context of this study, the Guerrieri et al. results suggest that neighborhood gentrification may be a confounding variable: advantageous home price performance in neighborhoods close to MARTA

[1] With the exception of possible time or neighborhood-level fixed effects.

stations could be a manifestation of rising incomes in that neighborhood rather than an externality of transit proximity. On the other hand, neighborhoods near MARTA stations could be likely candidates for gentrification precisely *because* MARTA is nearby. We control for this gentrification effect and identify an interesting new insight that suggests an interaction between gentrified neighborhoods and proximity to MARTA rail.

2. Data

```
df_pts <- readRDS("../data/resiliency_data.rds")
df_pts <- df_pts[complete.cases(df_pts@data), ]
#df_pts <- df_pts[sample(1:nrow(df_pts), 200), ]

df <- tbl_df(df_pts@data)
```

The Fulton County, Georgia tax assessor's office provided data for this study in response to a public records request. The appraised or assessed value of every single-unit residential property in the county from 2002 through 2012 is available, as is basic information about the structure such as the year of construction, the number of rooms in the building, and the size of the property. Of particular note is the "effective age," which is the age of the structure discounted by the tax assessor to accommodate reconstructions, renovations, and installed amenities that may not be original to the home (e.g., air conditioning, indoor plumbing, etc.). We use only home values from Fulton County to avoid idiosyncratic appraisal methods between counties and to ensure that a common set of covariates is available.

We extend the data by appending several additional variables have to these records. As a measure of neighborhood wealth, we use the home's Census tract median income recorded in the 2010 American Community Survey (U.S. Census Bureau, 2010b). Similarly, we measure the neighborhood racial composition as the percent of white people residing in each Census tract (U.S. Census Bureau, 2010a). We develop a measure of gentrification as the percent increase in Census tract real median income from 2000 to 2010. To measure proximity to public transportation, we measure the Euclidean distance from each home to the nearest MARTA station. We also measure the Euclidean distance to the nearest freeway entrance point, as an indicator of highway accessibility.

Although we have price evaluations available for each home in each of the ten years from 2002-2013, panel regression methods are inadequate, as the independent variables do not change for almost all homes over the time period in question. In particular, the MARTA rail network remained constant. It is therefore necessary to develop metrics by which the price performance of the home over time can be characterized.

2.1. *Price Performance*

Good investments express a number of characteristics. First, they have a positive net value: in a given period t, the expected value V of an investment i should be higher than its initial value,

$$\mathrm{E}\left[V_i(t)\right] > V_i(0) \tag{1}$$

with better investments having higher expected values. Additionally, good investments should have positive growth: the average proportional change in value is positive,

$$\mathrm{E}\left[\Delta V_i(t)\right] = \mathrm{E}\left[\frac{V_i(t+1) - V_i(t)}{V_i(t)}\right] > 0 \tag{2}$$

with better investments having higher expected growth rates. Finally, good investments have predictable growth: the variation in the growth rate is minimized,

$$\mathrm{Var}\left[\Delta V_i(t)\right] = 0. \tag{3}$$

These three metrics — mean value, mean growth rate, and standard deviation of the growth rate — were calculated for each of the homes in the dataset. The three functions used indexed values $V(t)/V(0), t \in 0, \ldots, 9$ to cancel the effect of initial price. Full descriptive statistics of all variables are given in Table 1.

```r
descriptive_stats <- df %>%
  # variables to show
  transmute(
    aacres = acres, brooms = rooms, cage = age,
    dincome = income_10,
    ewhite = pct_white,
    figrowth = income_grt,
    hmarta = marta_dist / 5280,
    iintchg = intchg_dist / 5280,
    jval = mean_val, kgrt = mean_grt, lsgrt = sdev_grt
  ) %>%
  # summary statistics
  summarise_each(
    funs(
      mean = mean(., na.rm = TRUE),
      median = median(., na.rm = TRUE),
      sd = sd(., na.rm = TRUE),
      first = quantile(., probs = 0.01, na.rm = TRUE),
```

```r
      nninth = quantile(., probs = 0.99, na.rm = TRUE)
    )
  ) %>%
  gather(key, value) %>%
  separate(key, c("variable", "stat")) %>%
  spread(stat, value, fill = NA) %>%
  select(variable, mean, median, sd, first, nninth)

## Warning: attributes are not identical across measure variables; they will be dropped

descriptive_stats$variable <- c(
  "Property acres", "Total rooms", "Effective home age",
  "Census tract median income", "Percent white", "Neighborhood income growth",
  "Distance to MARTA (miles)", "Distance to freeway interchange",
  "Mean indexed value", "Mean growth rate", "Standard deviation in growth rate"
)
```

[Table 1 about here.]

3. Spatial Analysis

We wish to describe a home's value performance as a function of its attributes, including proximity to a MARTA rail station. Following on the presentation of Macfarlane et al. (2015), it is necessary to control for spatial dependence, correlation, and endogenous missing variables. We do this with a *spatial Durbin model* (SDM),

$$\boldsymbol{y} = \rho W \boldsymbol{y} + X\boldsymbol{\beta} + WX\boldsymbol{\gamma} + \boldsymbol{\epsilon} \tag{4}$$

where $\boldsymbol{y}$ may be one of the three characterization functions. Other model elements include X, an $n \times p$ matrix where n is the number of homes and p is the number covariate attributes; $\boldsymbol{\epsilon}$, a stochastic error component assumed to have an independent and identical normal distribution; and W, an $n \times n$ matrix of weights mapping the spatial relationship between all i, j pairs $\{i, j\} \in 1, \ldots, n$. The model parameters $\rho, \boldsymbol{\beta}, \boldsymbol{\gamma}$ are estimated by maximum likelihood (ML). The average marginal direct, indirect, and total effects of a covariate $\boldsymbol{x}_k, k \in 1, \ldots, p$ in the SDM are

$$\begin{aligned} M(k)_{\text{direct}} &= n^{-1} tr((I - \rho W)^{-1}(I\beta_k + W\gamma_k)) \\ M(k)_{\text{total}} &= n^{-1} \boldsymbol{\iota}'(I - \rho W)^{-1}(I\beta_k + W\gamma_k)\boldsymbol{\iota} \\ M(k)_{\text{indirect}} &= M(k)_{\text{total}} - M(k)_{\text{direct}} \end{aligned} \tag{5}$$

where $\boldsymbol{\iota}$ is a vector of ones of length n. These effects and their empirical t-statistics can be obtained through a Monte Carlo simulation. We estimate model parameters and effects using maximum likelihood (ML) routines included in the `spdep` package for R (Bivand, 2013). It should be noted that the log-likelihood function for the SDM includes a log-determinant term $\ln|I - \rho W|$ that must be evaluated at every iteration of the maximization algorithm. This is a computationally expensive process of an order increasing with n. For this reason, we restrict our analysis to a random sample of $n = 5000$ observations.

In this study, we use a weights matrix W that considers the 50 nearest observations as neighbors, with a link weighted by the inverse distance between the observations. This weighting scheme helps to accommodate the changing density across the study area, and was found to have a good fit in terms of model likelihood relative to other matrices.[2] Figure 2 shows the neighborhood density in W by comparing the average distance between a home and all of its neighbors throughout the study area. Homes near to the region's core (and to other cities in the north part of the county) have a very low average neighbor distance, compared to suburban and exurban homes.

```
# find 50 nearest neighbors
nb <- knn2nb(knearneigh(df_pts, k=50))

## Warning in knearneigh(df_pts, k = 50): knearneigh: identical points found

# inverse distance between neighbors, forced to non-zero by adding fifty feet,
# and converted to miles
weights <- lapply(nbdists(nb, df_pts), function(x) 1 / ((x + 50) / 5280) )
W <- nb2listw(nb, glist = weights, style = "W")


# trace
trace <- trW(as(as_dgRMatrix_listw(W), "CsparseMatrix"))
```

In all of our models, we use a log-log specification, applying logarithmic transformations to each $\boldsymbol{y}$ and to the $\boldsymbol{x}$ variables representing property size, home age, median income, distance to rail, and distance to a freeway entrance. In this case, the estimated parameters express a constant elasticity, aiding interpretation; also, the ML algorithm converges more easily if the parameter estimates are on the same scale.

```
df_wgs <- spTransform(df_pts, CRSobj = WGS84) %>% coordinates(.)


df <- df %>%
```

[2] Details of the matrix selection process are available from the authors.

```r
  mutate(
    long = df_wgs[, 1],
    lat  = df_wgs[, 2],
    mean_neighbors = unlist(lapply(weights, function(x) mean(1 / x)))
  )
```

[Figure 2 about here.]

3.1. Results

```r
# set up model formulae
f_house <- formula(
  yval ~ log(acres) + rooms + log(age) + log(income_10) + pct_white
)

f_access <- update(f_house, . ~ . + log(marta_dist) + log(intchg_dist))

f_gentrify <- update(f_access, . ~ . + log(income_10))

f_interact <- update(
  f_house, . ~ . + income_grt * log(marta_dist) + log(intchg_dist)
)

formulae <- list(f_access, f_gentrify, f_interact)
names(formulae) <- c("access", "growth", "interacted")
dep_vars <- c("mean_val", "mean_grt", "sdev_grt")
```

```r
spatial_models = list()
for(i in names(formulae)){
  spatial_models[[i]] <- list()
  for(j in dep_vars){
    # set the right y value.
    df <- df %>% mutate_("yval" = as.name(j))
    if(j == "mean_val"){
```

```r
      df <- df %>% mutate(yval = log(yval))
    }
    spatial_models[[i]][[j]] <- lagsarlm(
      formulae[[i]],  data = df, listw = W, type = "mixed")
  }
}
```

We estimated SDM models of mean value, growth rate, and growth volatility with three sets of covariates. This results in a total of nine models. The simulated effects for each model and associated significance statistics are given in Tables 2 and 3; whereas Table 2 presents all three effects types for the mean value models, Table 3 shows only the total effects for the growth rate and volatility models.

```r
mean_impacts <- lapply(spatial_models, function(x)
  summary(impacts(x$mean_val, tr = trace, R=1000))
  )
```

[Table 2 about here.]

Beginning with the mean value models in Table 2, we can identify many of the trends previously observed by Macfarlane et al. (2015). Take the effective age, for instance: the expected mean value decreases with the age of a home, but increases with the weighted age of its neighbors. Intuitively, it is best to own a new home in an old (and presumably established or elite) neighborhood. The racial composition of a home's Census tract has no distinguishable direct effect, but rather a strong indirect effect with the expected value increasing with the share of white households. As the income of a home's neighbors increases, the expected value of a home *decreases*. This unintuitive result could be explained first by the fact that the initial value of a home has been normalized out of the equation, and second by the observations of Anderson and Beracha (2010) that home prices in wealthier neighborhoods are more sensitive to fluctuations in capital markets. Perhaps wealthier people lost more of their assets in the financial crisis, reducing demand (and subsequently prices) in wealthier neighborhoods, all else equal

```r
test1 <- lrtest(spatial_models$growth$mean_grt,
                spatial_models$interacted$mean_grt)
test2 <- lrtest(spatial_models$growth$mean_val,
                spatial_models$interacted$mean_val)
```

The effect of interest is the log(Distance to Rail) parameter, and in particular the total effect. In the "Access" model the estimated parameter of -0.011 is weakly significant, but has the hypothesized sign: all else equal, a home's average value decreases as the distance between the home and the rail station increases. The "Gentrify" model explores the possibility that the observed effect is not due to transit proximity *per se*, but rather neighborhood gentrification. Indeed, neighborhood income growth over the period is strongly correlated with an increase in home values, consistent with what would be expected through gentrification. The final model, "Interaction," includes an interaction term for the combined effects of income growth and rail proximity. This final model rejects the Gentrify model in a likelihood ratio test (p-value 0). In this model, the effect of income growth on its own becomes more pronounced, and the interaction term is strongly significant with a value of -0.059. Given two identical homes located in identical neighborhoods that are both growing in median income, the home located closer to transit will have a higher expected value across the study period.

```
growth_impacts <- lapply(spatial_models, function(x)
  summary(impacts(x$mean_grt, tr = trace, R=1000))
  )

volatility_impacts <- lapply(spatial_models, function(x)
  summary(impacts(x$sdev_grt, tr = trace, R=1000))
  )
```

[Table 3 about here.]

Table 3 presents the total effects for the three model specifications with the other dependent variables: the mean and the standard deviation of the growth rate over the period. Speaking first of the mean growth rate model, very few of the covariates are significant and show little change across the specifications. However, the interaction of income growth and rail transit proximity is significant, and as before, shows the expected sign. As income growth in a neighborhood increases, the average growth rate of home prices in that neighborhood will rise; as the distance between those homes and a MARTA station decreases, the growth rate will rise further.

In the case of the models estimating the relationship between home and neighborhood attributes and the standard deviation of the growth rate, the interaction between income growth and MARTA proximity does not significantly improve model likelihood. The effect of MARTA proximity on price volatility is weakly significant, though its negative sign indicates that homes prices closer to MARTA stations are *more volatile* than homes further away.

4. Latent Class Analysis

While the previous univariate analysis is informative, home value performance is by nature a multivariate problem. We can examine the joint performance of mean value, mean growth rate, and growth rate volatility with a multivariate finite mixture model,

$$H(Y|X, w, \psi) = \sum_{k=1}^{K} \pi_k(w, \alpha_k) \prod_{d=1}^{D} f_{kd}(Y_d|X_d, \theta_{kd}) \tag{6}$$

where the mixture density $H()$ is a function of a multivariate D-dimensioned response Y conditioned on the predictor variables X, concomitant variables w and model parameters $\psi = \{\alpha, \theta\}$. The concomitant parameters α define the probability π of each observation belonging in latent class k based on the values of w. The membership function π is a discrete response model, a multinomial logit model in our case. The relationship between the predictor variables X and each dependent variable Y_d is defined by the parameters θ_{kd} of the mixture function f_{kd}, which in our case is a Gaussian linear regression. For this initial analysis we restrict the number of classes to two, $k = 2$. We estimate the model using an expectation-maximization algorithm included in the `flexmix` package for R Leisch (2004).

This analysis has three goals: (1) identify whether the Fulton county housing market should be considered as more than one distinct market based on multidimensional value performance, (2) determine the variables that lead to inclusion in a more "favorable" market, and (3) determine if the markets show different relationships between the covariates and their outcomes.

```
interacted_flex <- flexmix(
  f_interact[-2], #rhs only
  concomitant = FLXPmultinom(f_interact[-2]),
  model = list(FLXMRglm(mean_val ~ .),
               FLXMRglm(mean_grt ~ .),
               FLXMRglm(sdev_grt ~ .)),
  data = df, k=2)

#make sure the smaller cluster is number one.
cluster_counts <- table(interacted_flex@cluster)
if(cluster_counts[1] > cluster_counts[2]){
  swap_cluster = TRUE
} else {
  swap_cluster = FALSE
}
```

```r
interacted_refit <- refit(interacted_flex)
df$class <- clusters(interacted_flex)
if(swap_cluster){ df$class <- abs(df$class - 3) }
df$class <- factor(df$class, labels = c("Class 1", "Class 2"))
```

We estimated four models, though we only present the model with likelihood sufficient to reject the other three. The first is an intercepts-only model with $\pi_k = \Lambda(\boldsymbol{\iota}\alpha_k)$ and $f_{kd} = \boldsymbol{\iota}\theta_{kd}$. The second is a concomitants-only class membership model using the set of covariates from Section 3 (including the interaction effect of income growth and transit proximity); thus $\pi_k = \Lambda(X\alpha_k)$ and $f_{kd} = \boldsymbol{\iota}\theta_{kd}$. The third model is the complement to the second, with the set of covariates used in the response functions; $\pi_k = \Lambda(\boldsymbol{\iota}\alpha_k)$ and $f_{kd} = X\theta_{kd}$. The fourth model includes a full set of covariates in the class membership function as well as the response functions, $\pi_k = \Lambda(X\alpha_k)$ and $f_{kd} = X\theta_{kd}$. This final model maximizes the model likelihood, and is therefore the model we present.

The model results allow us to reject the null hypothesis that $k = 1$, indicating that there are at least two latent classes of home price performance in the Atlanta region. The model predicts that most homes belong to Class 2 (91.7%). Figure 3 shows the price performance for a random sample of homes drawn from the model's membership assignments. The difference in trends between the two classes is immediately apparent. Homes in Class 1 experienced modest growth in prices until 2009, and then collapsed in value to a varying degree. Homes in Class 2, conversely, gained a substantial amount of value in the middle part of the period, and the average in the class remains higher than its initial value. Homes in Class 2 are likely to have a higher average value, a higher average growth rate, but also greater variance in the growth rate. Based on the analysis in the previous section, this cursory observation leads us to expect homes in Class 2 will be closer to MARTA stations.

[Figure 3 about here.]

Indeed, it appears that this is the case. Figure 4 compares the spatial density distribution for homes in Class 1, Class 2, and the full dataset. Class 1 is distributed throughout the county, but Class 2 is concentrated near the city center, where MARTA accessibility is high. The effect of proximity to MARTA on membership in Class 2 is further confirmed by the estimated parameters for the concomitant model given in Table 4. In this model positive estimates indicate a home is more likely to belong in Class 2 than Class 1 as the associated concomitant variable increases. With the exception of the distance to a freeway entrance, each of the estimates is highly significant. Membership in Class 2 is more likely as the property size decreases, the number of rooms in the home increases, the age of a home decreases, the median neighborhood income increases, and as the percent of white homes decreases. Also, and more relevant to our study, membership in Class 2 is less likely as the distance between a home and the nearest MARTA station increases.

The effect of neighborhood income growth and its interaction with transit proximity is more complicated than in the univariate SDM analysis. The negative coefficient estimates of the income growth and distance to rail parameters indicate that as neighborhood income growth or the distance to rail increases, homes are more likely to be in Class 2. But the significant and negative interaction term implies that at higher levels of income growth, homes further from transit are *less* likely to be in Class 2.

[Figure 4 about here.]

[Table 4 about here.]

An important feature of latent class mixture models is that the parameters θ_{kd} are allowed to have different effects on the response for each class. Figure 5 presents the coefficient estimates for the response models visually to aid comparison.[3] As the confidence intervals for many estimates overlap, we cannot reject that the effects of most covariates are the same across classes. There are some notable exceptions, however. The age of a home has no effect on any of the response variables for homes in Class 1, but a significant negative effect on each of the response variables for homes in Class 2. Simply, older homes are less valuable and experience lower growth only if the home is in Class 2, though older homes also have a less volatile growth rate. Conversely, the percent of white people in a neighborhood is more significantly correlated with the response variables for homes in Class 1. For these homes, their value and growth rate increase with the share of white neighbors, and variance in the growth rate decreases. The effects of income growth also differ for each class, increasing value volatility in Class 2 and decreasing it in Class 1.

```r
interacted_table <- prettify_refit(interacted_refit, type = "model") %>%
  mutate(model = as.character(factor(model, labels = dep_vars)))

if(swap_cluster){
  interacted_table$class <- as.character(abs(as.numeric(interacted_table$class) - 3))
}

sdm_table <- rbind_all(list(
  prettify_impacts(mean_impacts$interacted, "total", "plot", "mean_val"),
  prettify_impacts(growth_impacts$interacted, "total", "plot", "mean_grt"),
  prettify_impacts(volatility_impacts$interacted, "total", "plot", "sdev_grt")
))
```

[3] In this figure, the estimates and confidence intervals on the “Income Growth” variable have been scaled by 0.10 to improve clarity. Also, this figure includes the impacts from the spatial models presented previously for comparison.

```r
interacted_table <- rbind_all(list(interacted_table, sdm_table)) %>%
  mutate(
    model = factor(model, labels = c("Mean growth rate", "Mean value",
                                     "Std. dev. in growth rate")),
    beta = ifelse(variable == "income_grt", beta * 0.1, beta)
    )
```

In terms of the correlation between transit infrastructure and the response variables, the evidence is mixed. For homes in Class 1, proximity to MARTA rail is significantly correlated with higher average values, higher average growth rates, and lower growth volatility. But none of these correlations is significant for homes in Class 2. The same is true for freeway access, though the effects are smaller. But the interaction term with income growth is significant in all the response models, suggesting that in gentrifying neighborhoods closer to transit, mean values and mean growth rates are higher. But for volatility, the signs are reversed again.

[Figure 5 about here.]

5. Interpretations and Future Directions

The empirical results in general sustain the theories presented in Section 1.1. As the model of Glaeser et al. (2008) predicted, homes near MARTA have higher expected values but also higher volatility. And as the model of Guerrieri et al. (2013) predicted, homes in neighborhoods with increasing average incomes showed higher growth rates than homes in other neighborhoods; there was some additional effect for these neighborhoods if they were closer to transit stations. The significance of the interaction between income growth and distance to MARTA highlights that income growth in a Census tract can occur by two very different means: it could be a result of typical gentrification processes, with wealthy people displacing the poor and/or redeveloping former industrial zones; but income growth in a tract can also result from urbanization or suburban sprawl, as people move on to previous undeveloped land. Our analysis results suggest that it is the first process that leads to higher home values.

The two economic theories are complementary in that the Glaeser et al. model does not explain why markets may be inelastic, and the Guerrieri et al. model is agnostic to the cause of the demand shock. But the Guerrieri et al. model does not allow that the demand shock is irrational or even temporary, and thus provides no prediction on *where* prices will decline. Our empirical results, particularly the latent class membership analysis in Figure 3 and Table 4, suggest that prices increase primarily in the city center but decrease globally. Though the model coefficients suggest that transit proximity plays a role in this, we cannot rule out that there may be missing or endogenous variables for which MARTA is merely an indicator, such as population density or a loosely defined "urbanness."

The univariate analysis presented in Section 3 controls these missing variables and shows many of the same results, but without the subtleties of the multivariate, latent class analysis in Section 4. As an illustration of the differences between the models, consider the effects of income growth and rail proximity on growth rate volatility. In the spatial models none of the three coefficients is significant, but in the latent class model four of the six coefficients are significant. It is impossible to know how much of this discrepancy results from spatial dependence or correlation in the latent class models, and how much results from aggregation bias in the spatial models that assign a single coefficient estimate to at least two separate categories.

A potential resolution to this discrepancy is to introduce spatial effects into the mixture model, and these effects may need to be introduced in both the membership model as well as the response models. Wall and Liu (2009) present a univariate latent class model that introduces spatially correlated errors into the membership model. Gelfand and Vounatsou (2003) present a (single class) multivariate mixture model for continuous response variables that incorporates spatial dependence. We have not seen spatial effects introduced at both levels in our initial explorations of the literature. Developing such a model would be an important methodological contribution, as well as a necessary step in fully understanding this problem.

References

Alonso, W., 1960. A theory of the urban land market. *Papers in Regional Science* 6, 149–157. doi:10.1111/j.1435-5597.1960.tb01710.x.

Anderson, C.W., Beracha, E., 2010. Home price sensitivity to capital market factors: Analysis of ZIP code data. *The Journal of Real Estate Research* 32, 161–185.

Bivand, R., 2013. spdep: Spatial Dependence, Weighting Schemes, Statistics, and Models. URL: http://cran.r-project.org/package=spdep.

Bowes, D.R., Ihlanfeldt, K.R., 2001. Identifying the impacts of rail transit stations on residential property values. *Journal of Urban Economics* 50, 1–25. doi:10.1006/juec.2001.2214.

Debrezion, G., Pels, E., Rietveld, P., 2007. The impact of railway stations on residential and commercial property value: a meta-analysis. *The Journal of Real Estate Finance and Economics* 35, 161–180. doi:10.1007/s11146-007-9032-z.

Ewing, R., Hamidi, S., Absetz, S., Anderson, G., Berrigan, D., Chester, C., Dodds, A., Preuss, I., Tatalovich, Z., 2014. *Measuring Sprawl 2014*. Technical Report 4. Smart Growth America. Washington, D.C. URL: http://www.ncbi.nlm.nih.gov/pubmed/24942918, doi:10.1093/geronb/gbu023.

Gelfand, A.E., Vounatsou, P., 2003. Proper multivariate conditional autoregressive models for spatial data analysis. *Biostatistics* 4, 11–25. doi:10.1093/biostatistics/4.1.11.

Glaeser, E.L., Gyourko, J., Saiz, A., 2008. Housing supply and housing bubbles. *Journal of Urban Economics* 64, 198–217. doi:10.1016/j.jue.2008.07.007.

Grass, R.G., 1992. The estimation of residential property values around transit station sites in Washington, D.C. *Journal of Economics & Finance* 16, 139–146.

Guerrieri, V., Hartley, D., Hurst, E., 2013. Endogenous gentrification and housing price dynamics. *Journal of Public Economics* 100, 45–60. doi:10.1016/j.jpubeco.2013.02.001.

Iacono, M., Levinson, D., 2011. Location, regional accessibility, and price effects: evidence from home sales in Hennepin County, Minnesota. *Transportation Research Record* 2245, 87–94. doi:10.3141/2245-11.

Leisch, F., 2004. FlexMix : A general framework for finite mixture models and latent class regression in R. *Journal of Statistical Software* 11, 1–18.

Lewis-Workman, S., Brod, D., 1997. Measuring the neighborhood benefits of rail transit accessibility. *Transportation Research Record* 1576, 147–153.

Macfarlane, G.S., Garrow, L.A., Moreno-Cruz, J., 2015. Do Atlanta residents value MARTA? Selecting an autoregressive model to recover willingness to pay. *Forthcoming, Transportation Research Part A: Policy and Practice* .

McMillen, D.P., McDonald, J., 2004. Reaction of house prices to a new rapid transit line: Chicago's Midway Line, 1983-1999. *Real Estate Economics* 32, 463–486.

Pivo, G., 2013. The effect of transportation, location, and affordability related sustainability features on mortgage default prediction and risk in multifamily rental housing. URL: http://www.fanniemae.com/resources/file/aboutus/pdf/hoytpivo_mfhousing_sustainability.pdf.

Pollack, S., Bluestone, B., Billingham, C., 2010. *Maintaining diversity in Americas transit-rich neighborhoods: Tools for equitable neighborhood change*. Technical Report. Dukakis Center for Urban and Regional Policy. URL: http://iris.lib.neu.edu/dukakis_pubs/3/.

Smith, J.J., Gihring, T.A., 2006. Financing transit systems through value capture. *American Journal of Economics and Sociology* 65, 751–786. doi:10.1111/j.1536-7150.2006.00474.x.

Standard & Poors, 2013. Case-Shiller Home Price Index. URL: http://us.spindices.com/index-family/real-estate/sp-case-shiller.

U.S. Census Bureau, 2010a. *B02001. Race: 2010 ACS 5-year estimates.* Technical Report.

U.S. Census Bureau, 2010b. *B19013. Median household income in the past 12 months: 2010 ACS 5-year estimates.* Technical Report. URL: http://factfinder2.census.gov/.

Wall, M.M., Liu, X., 2009. Spatial Latent Class Analysis Model for Spatially Distributed Multivariate Binary Data. *Computational statistics & data analysis* 53, 3057–3069. doi:10.1016/j.csda.2008.07.037.

List of Figures

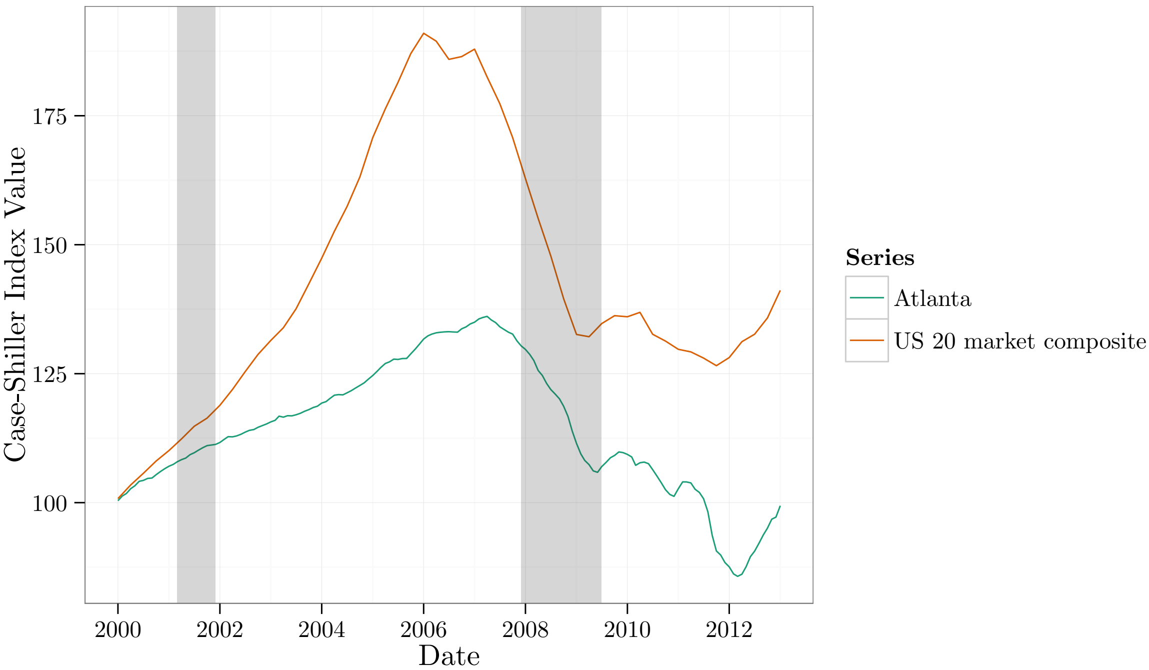

Figure 1: Case-Shiller home price index, seasonally adjusted.

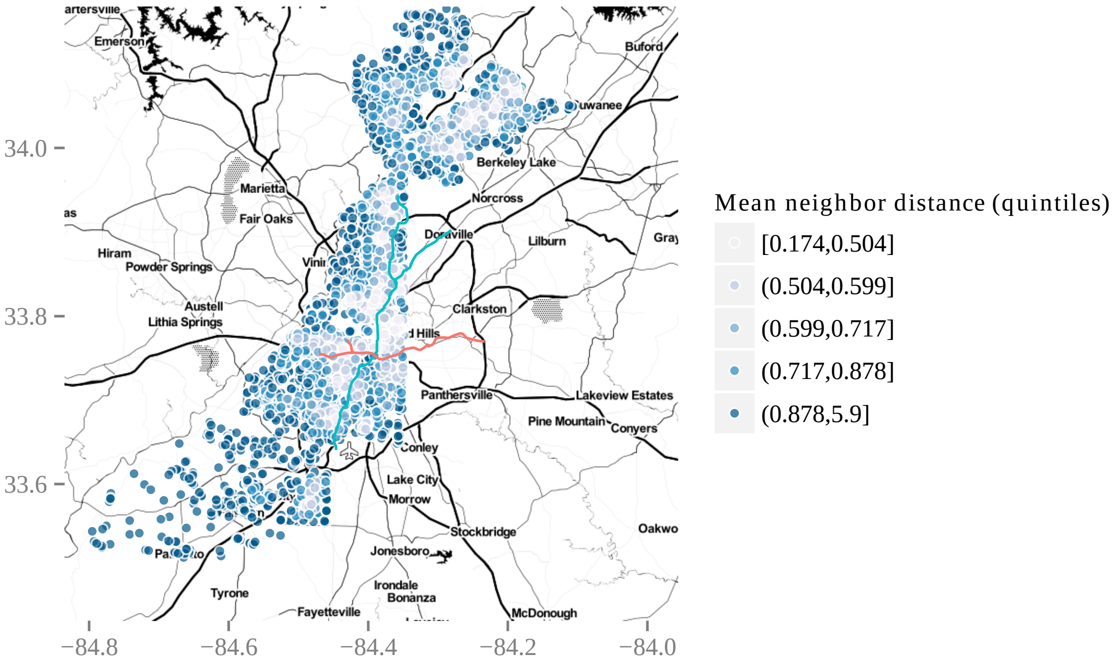

Figure 2: Observations in spatial models, by average distance of 50 nearest neighbors.

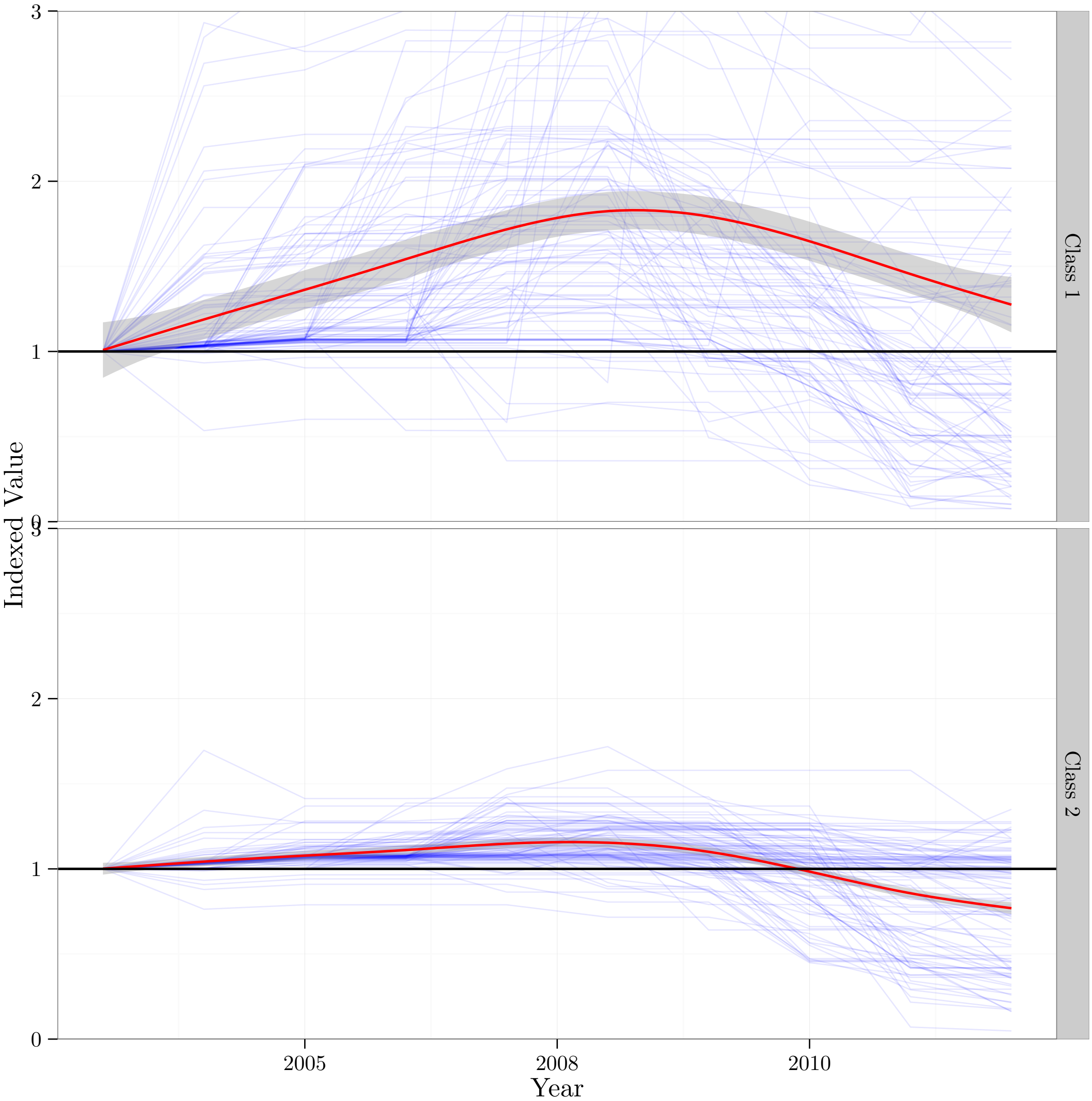

Figure 3: Value performance for a sample of 100 homes in each latent class.

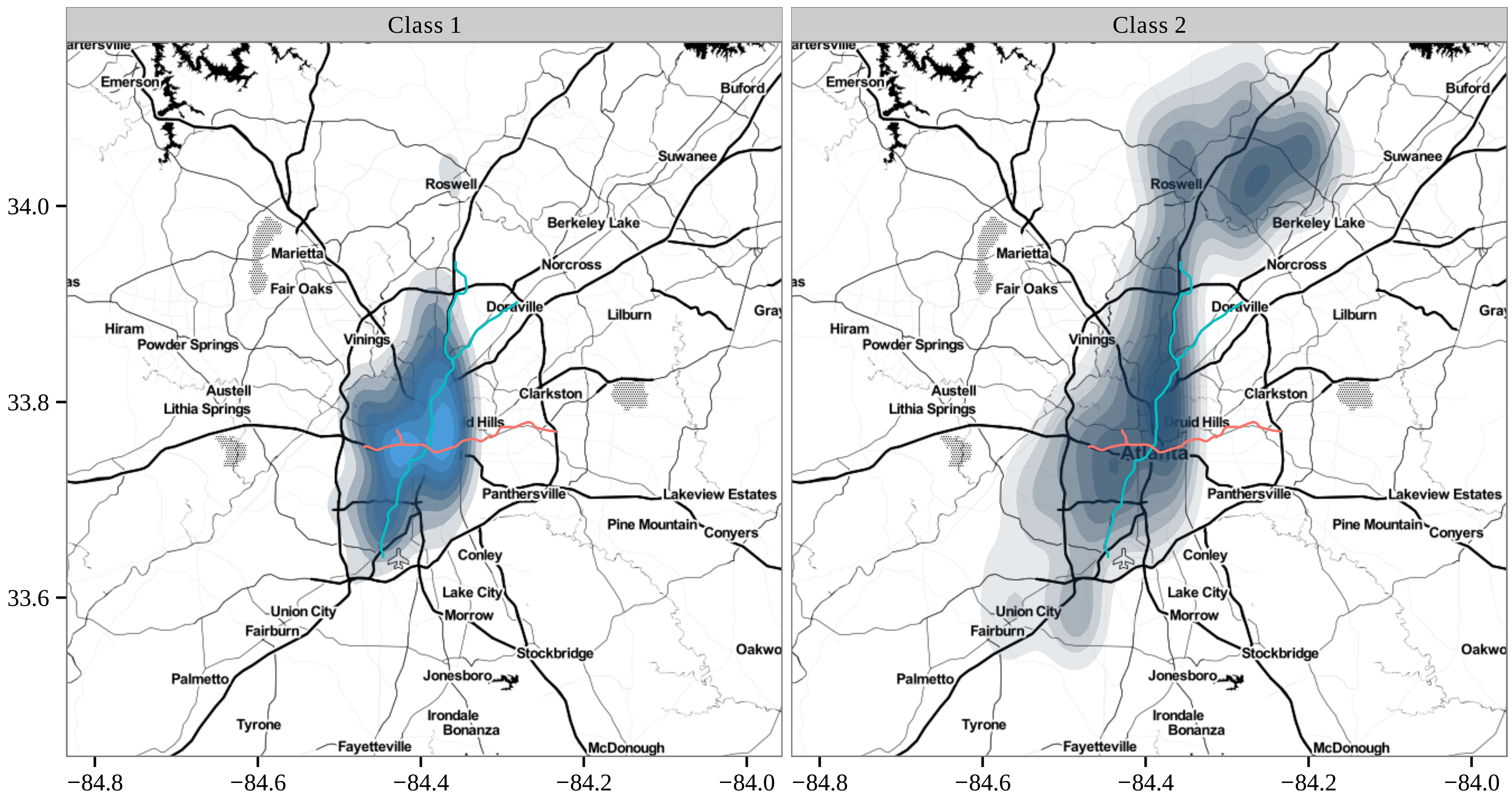

Figure 4: Relative spatial density of homes in each latent class.

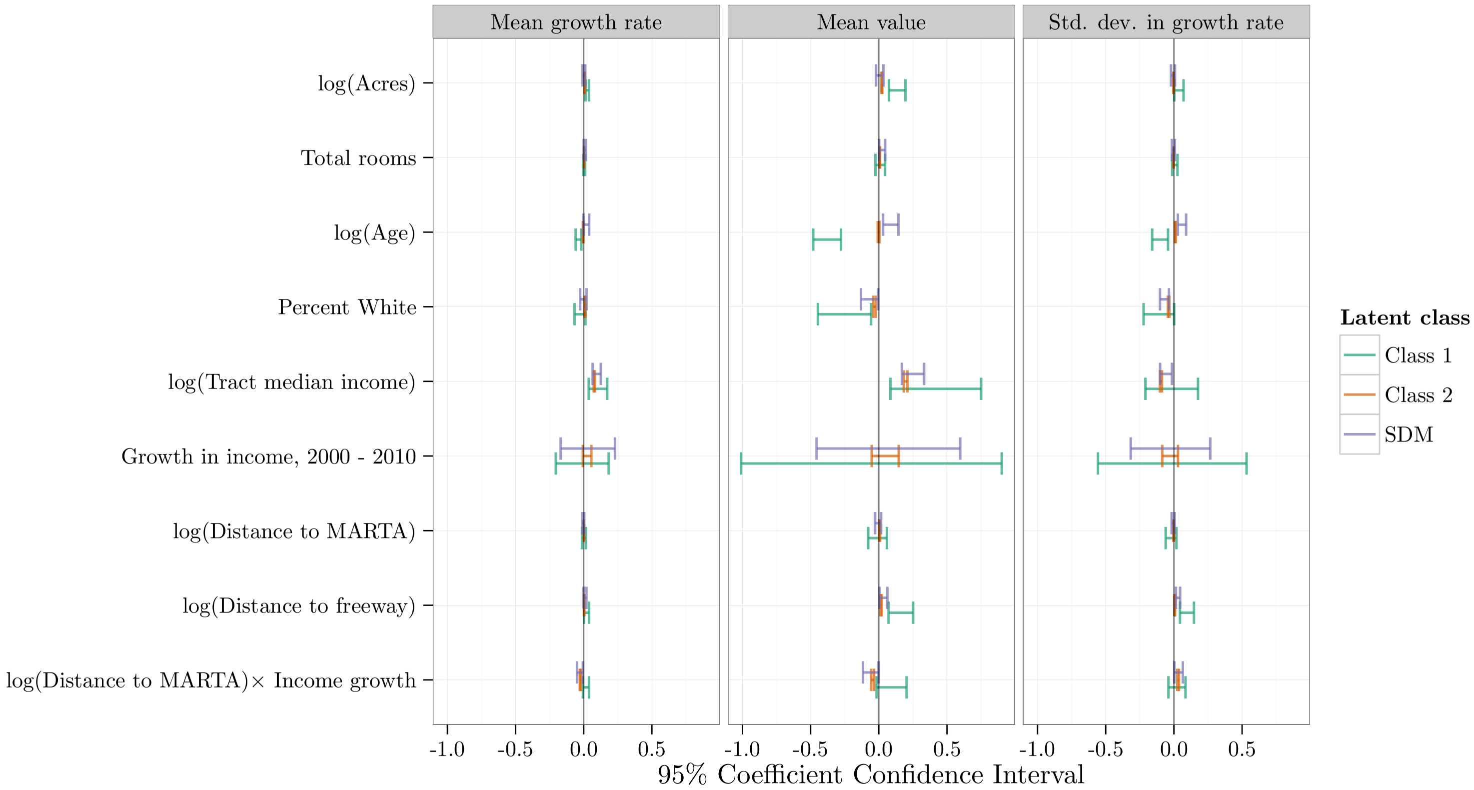

Figure 5: Latent class covariate estimates and confidence intervals.

List of Tables

Table 1: Descriptive statistics of model variables

	Mean	Median	Std. Dev.	1%	99%
Independent variables					
Property acres	0.554	0.290	1.745	0.015	5.170
Total rooms	6.683	6.000	1.844	3.000	12.000
Effective home age	37.203	32.000	19.721	12.000	94.000
Census tract median income	74785.313	70787.000	41360.177	16321.000	176818.000
Percent white	0.512	0.635	0.344	0.001	0.957
Neighborhood income growth	-0.108	-0.175	0.263	-0.492	0.740
Distance to MARTA (miles)	4.710	3.198	4.132	0.215	15.004
Distance to freeway interchange	1.846	1.371	1.463	0.177	6.500
Dependent variables					
Mean indexed value	1.056	1.039	0.232	0.781	1.811
Mean growth rate	-0.022	-0.010	0.060	-0.130	0.127
Standard deviation in growth rate	0.134	0.104	0.132	0.013	0.572

Table 2: Effects of covariates on mean home value.

	Access		Gentrify		Interaction	
Covariates	$M(k)$	t-stat	$M(k)$	t-stat	$M(k)$	t-stat
Direct Effects						
log(Acres)	0.04^{***}	14.87	0.04^{***}	14.88	0.04^{***}	14.31
Total rooms	-0.002	-1.32	-0.002	-1.32	-0.001	-0.94
log(Age)	-0.069^{***}	-14.07	-0.069^{***}	-13.37	-0.069^{***}	-13.35
Percent White	$0.081^{\dagger}$	1.86	$0.079^{\dagger}$	1.87	0.057	1.27
log(Tract median income)	-0.027^{*}	-1.99	$-0.026^{\dagger}$	-1.85	-0.054^{**}	-3.28
Growth in income, 2000 - 2010					-0.162	-1.1
log(Distance to MARTA)	0.031^{*}	2.1	0.032^{*}	2.05	0.031^{*}	2.03
log(Distance to freeway)	0.004	0.38	0.005	0.48	0.001	0.12
log(Distance to MARTA)$\times$ Income growth					0.024	1.5
Indirect Effects						
log(Acres)	$-0.026^{\dagger}$	-1.78	$-0.026^{\dagger}$	-1.66	-0.034^{*}	-2.4
Total rooms	0.009	0.76	0.008	0.7	0.026^{*}	2.35
log(Age)	0.145^{***}	4.67	0.146^{***}	4.74	0.157^{***}	5.42
Percent White	0.159^{*}	2.4	0.16^{*}	2.36	0.194^{**}	2.99
log(Tract median income)	0.011	0.29	0.009	0.25	-0.014	-0.37
Growth in income, 2000 - 2010					0.866^{**}	2.58
log(Distance to MARTA)	-0.042^{*}	-2.12	-0.042^{*}	-2.1	$-0.037^{\dagger}$	-1.81
log(Distance to freeway)	0.026	1.16	0.024	1.13	0.033	1.55
log(Distance to MARTA)$\times$ Income growth					-0.083^{*}	-2.28
Total Effects						
log(Acres)	0.014	0.99	0.014	0.95	0.007	0.5
Total rooms	0.007	0.59	0.006	0.53	0.024^{*}	2.2
log(Age)	0.076^{*}	2.48	0.076^{*}	2.5	0.088^{**}	3.06
Percent White	0.24^{***}	5.35	0.24^{***}	5.18	0.251^{***}	6.06
log(Tract median income)	-0.016	-0.52	-0.017	-0.52	-0.068^{*}	-2.13
Growth in income, 2000 - 2010					0.704^{**}	2.63
log(Distance to MARTA)	-0.011	-0.99	-0.01	-0.95	-0.005	-0.48
log(Distance to freeway)	$0.03^{\dagger}$	1.91	$0.029^{\dagger}$	1.87	0.034^{*}	2.32
log(Distance to MARTA)$\times$ Income growth					-0.059^{*}	-2.05
Model Log-likelihood	$3,047$		$3,047$		$3,066$	

† significant at $p < .10$; $^{*}p < .05$; $^{**}p < .01$; $^{***}p < .001$

Table 3: Total effects of covariates on home performance.

Covariates	Access		Gentrify		Interaction	
	$M(k)_{\text{Total}}$	t-stat	$M(k)_{\text{Total}}$	t-stat	$M(k)_{\text{Total}}$	t-stat
Mean Growth Rate						
log(Acres)	0.003	0.62	0.003	0.64	0.002	0.32
Total rooms	0.002	0.62	0.002	0.56	$0.008^{\dagger}$	1.88
log(Age)	0.013	1.18	0.014	1.17	$0.019^{\dagger}$	1.79
Percent White	0.092^{***}	5.62	0.091^{***}	5.94	0.095^{***}	6.26
log(Tract median income)	0.011	0.92	0.012	1.04	-0.003	-0.25
Growth in income, 2000 - 2010					0.3^{**}	2.96
log(Distance to MARTA)	-0.004	-1.1	-0.004	-1.05	-0.004	-1.02
log(Distance to freeway)	0.008	1.53	0.008	1.41	$0.009^{\dagger}$	1.76
log(Distance to MARTA)$\times$ Income growth					-0.028^{*}	-2.55
Model Log-likelihood	8,617		8,617		8,631	
Standard Deviation of Growth Rate						
log(Acres)	0	-0.02	0	0.05	-0.006	-0.87
Total rooms	-0.008	-1.39	-0.008	-1.41	-0.004	-0.68
log(Age)	0.061^{***}	3.8	0.061^{***}	3.82	0.059^{***}	3.86
Percent White	-0.052^{*}	-2.31	-0.051^{*}	-2.34	-0.058^{**}	-2.63
log(Tract median income)	-0.057^{***}	-3.42	-0.057^{***}	-3.61	-0.069^{***}	-4.17
Growth in income, 2000 - 2010					$-0.251^{\dagger}$	-1.69
log(Distance to MARTA)	-0.013^{*}	-2.33	-0.013^{*}	-2.39	-0.006	-1
log(Distance to freeway)	0.029^{***}	3.55	0.028^{***}	3.59	0.03^{***}	3.92
log(Distance to MARTA)$\times$ Income growth					0.034^{*}	2.15
Model Log-likelihood	3,909		3,909		3,918	

† significant at $p < .10$; $^{*}p < .05$; $^{**}p < .01$; $^{***}p < .001$

Table 4: Concomitant (class membership) model.

	Class 2	
Covariate	α	t-stat
(Intercept)	-25.587^{***}	-10.79
log(Acres)	-25.587^{***}	-6.31
Total rooms	-25.587^{*}	2.39
log(Age)	-25.587^{***}	3.64
Percent White	-25.587^{***}	8.74
log(Tract median income)	-25.587^{***}	-5.27
Growth in income, 2000 - 2010	-25.587	-1.05
log(Distance to MARTA)	-25.587^{***}	6.66
log(Distance to freeway)	-25.587	0.10
log(Distance to MARTA)$\times$ Income growth	-25.587	-0.18

** significant at $p < .01$; $^{***}p < .001$