A Semiparametric Investigation of Lower-Income Home Mortgage Purchases in the Secondary Mortgage Market[*]

January 2000

Dapeng Hu

Zell/Lurie Real Estate Center
The Wharton School
University of Pennsylvania
Email: hudapeng@wharton.upenn.edu

Abstract

The role that neighborhood characteristic plays in home mortgage lending in both primary and secondary mortgage markets has recently received increasing attention in housing finance research. However, the social-economic characteristics of a neighborhood are highly correlated with the distribution of loan applicants' credit risk in the neighborhood. Therefore, it remains difficult to separate the role of neighborhood characteristics from the effect of applicant's credit risk factors. Using a newly developed additive semiparametric regression technique, this paper investigates the spatial distribution of the lower income mortgage purchases in the secondary mortgage market while focusing on the non-linear effect of applicant's credit risk factors. Since semiparametric models do not impose any specification on borrowers credit risk factors, they avoid potential mis-specification problem and generate consistent estimations of the effect of neighborhood characteristics. Furthermore, the paper graphically presents the non-linear effects of borrower's credit risk factors such as income and LTV on the lower income mortgage purchases. It is the graphical representation of these non-linear components that provides a new and useful tool for analyzing mortgage risks.

JEL Classification: G21, R10, C14

[*] This paper extends an earlier research of Gyourko and Hu, entitled "Intra-metropolitan Distribution of GSEs Affordable Housing Mortgage Purchases," which was supported by a grant from the Department of Housing and Urban Development and the Research Sponsors Program of the Zell/Lurie Real Estate Center at Wharton. The author is grateful to Joe Gyourko's insightful input. The author also thanks Anthony Pennington-Cross, Amy Crews-Cutts, Frank Nothaft, Charles Calhoun, Theresa DiVenti, Wayne Passmore, Dick Voith, Peter Zorn, Yongheng Deng, and Todd Sinai for helpful comments.

1. Introduction

The role that neighborhood characteristic plays in home mortgage lending in both primary and secondary mortgage markets has recently received increasing attention in housing finance research. However, the social-economic characteristics of a neighborhood are highly correlated with the distribution of loan applicants' credit risk in the neighborhood. Therefore, it remains difficult to separate the role of neighborhood characteristics from the effect of applicant's credit risk factors. For example, in a recent study of the affordable home mortgage purchases of the two Government Sponsored Enterprises (GSEs)[1], Gyourko & Hu (1999) have shown that the center city's share of GSEs lower income purchases in a metropolitan area is materially smaller than the center city's share of lower income homebuyers. Is this because the credit risks of lower income homebuyers in the center city are generally higher than those in suburban, or because the GSEs (and/or primary lenders) are simply avoiding center cities? Correctly gauging applicants' credit risk factors is critical to the estimation of the effect of neighborhood characteristics; otherwise, the conclusion would be misleading.

Using a newly developed additive semiparametric regression technique, this paper investigates the spatial distribution of the GSEs lower income mortgage purchases, focusing on the non-linear effect of applicants' credit risk factors. Since semiparametric models do not impose any specification on borrower's credit risk factors, they avoid potential mis-specification problems, and generate consistent estimations of the effect of neighborhood characteristics. Furthermore, the paper graphically presents the non-linear nature of the effects of borrower's risk factors such as income and LTV. It is the graphical representation of these non-linear components that we feel provide a new and useful tool for analyzing mortgage risks.

[1] In exchange for their federal chartered status, the two GSEs, Fannie Mae and Freddie Mac, are required to help provide affordable mortgage loans to lower income families and distressed areas. Three affordable housing goals were promulgated by the Secretary of the Department of Housing and Urban Development (HUD): the Low- and Moderate-Income Goal, targeting all the lower-income families; the Geographically Targeted Goal targeting underserved areas; and the Special Affordable Goal targeting very low-income households. See HUD (1995).

A growing literature studies the role that neighborhood characteristic plays in mortgage performance and mortgage origination recently. Based on cross-tabulations, some earlier works had examined the associations between neighborhood income, racial component, and center city location with mortgage lending at the national aggregate level (Canner, 1995; MacDonald, 1996; Manchester, 1998). While strong associations between neighborhood characteristics and mortgage origination rate were identified, these studies typically did not control for the borrower's credit risk distribution. Van Order et al (1993) and Berkovec et al (1994, 1998) examined neighborhood and borrower race status on mortgage default rates; Anselin and Can (1998) examined the spatial effects on mortgage lending in Atlanta metropolitan area; Harrison (1999) reported the neighborhood effects on primary market acceptance/rejection rate; Calem and Watchter (1999) examined long-term delinquency in relation to neighborhood housing market conditions and borrower credit history. These studies found that neighborhood characteristics, in addition to borrower-specific risks, are significantly correlated with mortgage performance. Gyourko and Hu (1999) studied the spatial distribution issue of secondary market affordable housing liquidity and revealed a spatial mismatch[2] in a sense that the intra-metropolitan distribution of GSE affordable mortgage purchases is materially different from the distribution of the targeted potential homebuyers. As in all previous studies, linear or log-linear relations between borrower's risk factors and loan performance were typically specified in these studies.

The linearity assumption between borrower's credit risk factors and GSE mortgage purchases is unlikely to hold true. Lending guidelines and underwriting standards can not be accurately characterized as having smooth or linear relationship with borrower risks.[3] Applicants are typically grouped into risk categories and then assigned appropriate prices and underwriting standards. For example, conventional

[2] The term spatial mismatch here is obviously in a different context from that of Kain's hypothesis (Kain, 1962).

[3] Discontinuities in lending standards have grown out of the inability of lenders to accurately price or identify credit risks. For instance, a borrower with a 21% down payment is treated much differently than a borrower with a 19% down even though the credit risks may be very similar. These discontinuities also are inherent in the treatment at payment to income and credit history factors.

loans with LTV above 80% are typically required to buy mortgage insurance, which reduces the risk borne by the lender. As Canner and Passmore (1995) point out, the actual decision of whether or not a mortgage is originated rests upon the entity that ultimately bears the credit risk; and Private Mortgage Insurers (PMIs) bear the majority of the credit risk when a loan is insured. Therefore, the effect of LTV on the distribution of GSE purchases is very likely to be non-linear or even non-monotonic.

The effect of borrower's income on GSE purchases may not be either linear or log-linear. Typically, a borrower's income is an important measurement for credit risk; it is negatively related to payment-to-income ratio and positively related to the credit scores. However, GSEs have the obligation to fulfill their "special affordable" goal, which targets very low-income families with income below 60% of area median income (AMI) or below 80% of AMI if in a poor census tract. Therefore, the loan applicants who are eligible for this special affordable goal may have a higher chance of being picked up by GSEs.

The above arguments suggest a non-linear treatment of borrower's credit risk factors in modeling the distribution of GSE lower income purchases. Nonparametric and semiparametric approach provides a powerful tool for dealing with non-linearity and non-normal data distribution.[4] Using a newly developed additive semiparametric techniques (Linton and Neilsen, 1995; Fan, Hardle and Marmen, 1996), this paper examines the non-linear effects of borrower's income and LTV on the spatial distribution issue of the GSEs' lower income loan purchases. Because the borrower's risk factors are estimated nonparametrically and do not rely on a linear specification, the semiparametric approach also helps to overcome the multi-collinearity between borrower factors and neighborhood characteristics.

The semiparametric approach that we use is a partial linear regression (PLR) model which allows the unknown non-linear components to enter additively. It has an advantage over the typical semiparametric regression model developed by Robinson (1988) in that the semiparametric additive model

[4] Because results of highly non-linear models are very sensitive to the choice of the parametric form and the distribution of observable variables, nonparametric or semiparametric models often become preferable alternatives (see Barnett, Powell, & Tauchen, 1991).

allows for explicit estimation of the marginal effects of these non-linear components on the dependent variable, whereas the traditional semiparametric formulation treats the variables that enter the unknown non-linear part of the model as nuisance variables. A recent application of this method (Liu and Stengos, 1999) shows its power in revealing non-linearity in the data.

In addition, this paper controls for the effect of loan-type differences on the GSE purchases. Investor loans, for example, typically have a higher risk than owner's loans and may be correlated with certain neighborhood characteristics. The GSE Public Used Database and Home Mortgage Disclosure Act (HMDA) data are employed in a detailed spatial analysis of 20 large metropolitan areas; there areas account for 31.2% of the nation's households and 31.7% of GSE single-family mortgage purchases in 1996.

Contrary to linear model results, the PLR model found no consistent evidence that indicates central city neighborhoods being underserved after controlling for borrower's risk factors and loan type frequency in each tract. Racial and income status of a neighborhood, in addition to the borrower's risk factors and loan type factors, do contribute to the distribution of GSE lower income mortgage purchases. African-American concentrated tracts are more likely to be under-represented if in suburban locations; however, there is no indication that heavily minority areas in central city tracts are under-represented. GSE lower income mortgage purchases are quite high, relative to the number of lower income borrowers in rich neighborhoods.[5]

Finally, the non-linear relationship between borrower's risk factors and the GSE purchases is examined by graphical presentation of the nonparametric part of the PLR model. While the borrower's income is positively related to the GSE purchase rate, the PLR model reveals that there is a higher GSE purchase rate at the low end of income range, which may due to the effect of the Special Affordable Goal

[5] That said, our results do not, in and of themselves, imply that the mismatch is necessarily caused by GSEs' lending behavior and it would be inappropriate to infer conclusions regarding discrimination from this paper.

(Goal 3). The LTV effect is complicated and varies across metropolitan areas. In most areas, there is a negative correlation between tract mean LTV and the tract's GSE purchase rate when the tract mean LTV is below 80%. However, this negative correlation typically stops at around 80% in the tract mean LTV -- and it turns into a positive correlation in several metropolitan areas at the high end of LTV. A possible explanation for this result may rest on the mortgage insurance or other non-linearity in the mortgage lending and purchasing process.

2. Model and semiparametric specification

In this section, we derive a model showing that the spatial distribution of low-income loan purchases can be a function of neighborhood characteristics, borrowers credit risk, and loan type information, driven by lender's profit maximization behavior. Since our concern is ultimately how the implicit subsidies in the secondary mortgage market are distributed to low-income families, we do not distinguish the primary market lender and the secondary market purchasers in the model.[6]

House price and price appreciation are strongly related to neighborhood characteristics. In addition, spatial spillover effect may also play an important role in mortgage lending. For example, the deterioration of one property may cause the depreciation of other properties in its vicinity. Because spatial spillover effects are typically not fully capitalized in housing price or appraised housing value,[7] properties in certain types of neighborhoods may have higher appreciation rates than similar properties in other neighborhoods (Archer, et al 1996; Crone & Voith, 1999). In a distressed area, the decline in housing value is more severe, which increases the probability of negative equity positions and, therefore, a higher credit risk.

[6] Although one can argue that GSEs can influence the primary lenders' behavior (see the report of Buist et al, 1994), this paper does not intend to suggest that the GSEs are solely responsible for the spatial distribution of low-income loan purchases. By comparing the GSE lower-income purchases with the lower-income loans originated by primary lenders, we find no substantial difference from the data (see result in Appendix A2).

[7] The housing market inefficiency has been argued by Case and Shiller (1989) and relation of MSA characteristics with housing price appreciation has been captured by Clapps and Giaccoto (1994).

It is argued that information externality, for example the accuracy of appraisal price, may be related to housing market conditions (see Lang and Nakamura, 1993). This suggests an additional reason for the correlation of risk with neighborhood traits. The difference of appraisal price and contract sale price is an important factor in the residential mortgage lending. Empirical studies has shown that low appraised value is related to proxies for neighborhood quality (LaCour-Little and Green, 1997).

Asymmetric information may provide another reason that neighborhood traits play a role in lower-income mortgage lending or purchasing. Individual traits, such as income, have limited variation within the pool of goal-eligible borrowers, and sometimes are even unverifiable. Legal issues apart, lenders may want to use other factors (such as location) that correlated with credit risks. Previous studies have found that foreclosure rates tends to be higher in distressed or economically disadvantaged neighborhoods (see the review of Quercia & Stegman 1992; and Berkovec et al., 1994). Hence, lenders and GSE mortgage purchases rationally could use group traits (neighborhood characteristics) to help mitigate the costs associated with potential defaults.

The true credit risk $\bar{u}_{ij}$, for a loan application i with underlying property in tract j is unknown to lenders. Lenders perceive the risk based on applicant's credit risk factors (*Z*), neighborhood traits (*X*), and loan type information (*S*), as in equation (1)

$$(1) \qquad \tilde{u}_{ij} = E(\bar{m}_{ij} \mid Z_{ij}, S_{ij}, X_j) + \boldsymbol{h}_{ij},$$

where $\tilde{u}_{ij}$ is the perceived risk, $\boldsymbol{h}_{ij}$ is an error term, and *Z*, *X*, and *S* are vectors of variables.

When a loan defaults, we denote the cost as c_{ij}, which may also vary by individual property or neighborhood. The expected default cost (v_{ij}), given perceived default risk, then equals

(2) $$v_{ij} = \tilde{u}_{ij}\, c_{ij}\ .$$

Theoretically, the true risk premium, or the efficient price for default risk can be derived from option pricing models; if the market is efficient and complete, then the realized price should be the derived equilibrium price plus a white noise. However, the real market is not perfect and not necessarily complete and, it is difficult to price the risk in practice.[8] Absent the ability to easily price credit risks (in mortgage rate, points, or fees), a lender's profit maximizing problem boils down to one of minimizing expected default costs, assuming other costs (cost of fund and transaction cost) are invariant across loans. Therefore the v_{ij} becomes the criteria for GSEs to determine if a low-income application be picked up or not.

If we denote the marginal probability for any low-income mortgage application to be picked up by GSEs in metropolitan area c as $prob_c$, the conditional probability for a low-income application i in tract j to be picked up by GSEs, $prob(loan \mid i, j)$, is then a function of the expected default cost v_{ij} and $prob_c$,

(3) $$prob(loan \mid i, j) = g(v_{ij}) * prob_c\,,\ g \geq 0,\ \text{and}\ g' \leq 0\ .$$

The $prob_c$ depends upon loan supply and demand conditions in a metropolitan area, specifically,

(4) $$prob_c = \sum_{j=1}^{n} l_j \Big/ \sum_{j=1}^{n} h_j\ ,$$

[8] Legal or political considerations are factors here, in addition to there being relatively little heterogeneity among goal-eligible borrowers. GSEs benefit financially from their special status and they would put that benefit at risk if they were to perfectly price discriminate.

where l_j is the number of low-income loan purchases in tract j, h_j is the number of low-income applicants in tract j, and n is the total number of census tracts in a metropolitan area.

The expected number of loan purchases in tract j then equals

$$(5) \qquad E(l_j) = h_j E_i[\,prob(loan \,|\, i, j)] = h_j E_i[g\,(v_{ij})\frac{\sum_{j=1}^{n} l_j}{\sum_{j=1}^{n} h_j}] = h_j \frac{\sum_{j=1}^{n} l_j}{\sum_{j=1}^{n} h_j} E_i[g\,(v_{ij})]\,.$$

Rearranging equation (5) and substituting equation (2) into (5), we obtain

$$(6) \qquad \frac{E(l_j)}{h_j} \Bigg/ \frac{\sum_{j=1}^{n} l_j}{\sum_{j=1}^{n} h_j} = E_i[g\,(v_{ij})] = E_i\,[m(Z_{ij}\,, S_{ij}\,, X_j)].$$

The RHS of equation (6) is the purchase rate in tract j normalized by the MSA's average purchase rate, which is named as the normalized purchase rate (*NPR*), $NPR_j \equiv \frac{E(l_j)}{h_j} \Big/ \frac{\sum_{j=1}^{n} l_j}{\sum_{j=1}^{n} h_j}$. With a little rearrangement, *NPR* can also be rewritten as $\frac{E(l_j)}{\sum_{j=1}^{n} l_j} \Big/ \frac{h_j}{\sum_{j=1}^{n} h_j}$, which is a tract's share of low- income loans over the tract's share of goal-eligible applicants. If a tract's *NPR* is less than one, then the tract's share of loan purchases is less than its share of eligible applicants; if a tract's *NPR* is greater than one, then its share of loans purchased exceeds its share of goal-eligible borrowers. Therefore, *NPR* also measures how low-income loan purchases are spatially distributed relative to the distribution of lower income borrowers.

Assuming the linearity of the loan type variables in function m and assuming the distribution of borrower's risk factors Z can be represented by the means of these factors $\bar{Z}_j$, equation (6) becomes

(7) $$NPR_j = f(\bar{Z}_j, s_j, X_j)$$

where $\bar{Z}_j = E(Z_{ij})$. That is, the relative spatial distribution (or the normalized GSE purchase rate in each tract) is a function of neighborhood characteristics, the frequency of each loan type s_j, and the mean of the vector of borrower's traits $\bar{Z}_j$.

From equations (6) and (3), we know that $\frac{\partial NPR_j}{\partial v_{ij}} \leq 0$. Therefore for a profit maximizing agent, we have

(8) $$\frac{\partial NPR_j}{\partial x_j} \frac{\partial v_j}{\partial x_j} \leq 0.$$

where x is any variable of the neighborhood characteristics, borrower's risk factors, or loan type information. For example, if lenders view certain neighborhood characteristics as risky, we should observe a negative correlation between that character and the relative distribution of loan purchases.

We assume linearity of the neighborhood characteristics and loan type information, and the non-linearity of borrower's risk factors. We can have the following Partial Linear Regression (PLR) model specification for equation (7),

(9) $NPR_j = X_j^T \boldsymbol{b} + s_j^T g + g(\overline{Z}_j) + \boldsymbol{e}_j$ *, j=1,…,n,*

where b and g are vectors of unknown parameters on vector X_j and vector *s*, and g(.) is an unknown function.

Robinson (1988) provided a way of obtaining $\sqrt{n}-$consistent estimator of the parameter vector b and g, by concentrating out the influence of the nuisance variables Z's by conditioning on them. We use kernel methods (see Hardles, 1990) to estimate the conditional expectations $E(PR_j \mid Z_j)$, $E(X_j \mid Z_j)$, and $E(PR_j \mid Z_j)$. The kernel estimation is then used in the second stage of a two-step estimation procedure to estimate the parameter vector b and g. Comparing to OLS, this approach provides better control over the borrower's risk factors Z. However, because it also conceals the influence of the Z's in the regression function,[9] it does not allow us to see how the Z's work on the dependent variable.

To see the effects of the borrower's risk factors Z's on the spatial distribution *NPR*, we have to impose more structural restriction on model (10) by allowing an additive structure on the unknown components. That is the regression model can now be written as

(10) $NPR_j = X_j^T b + s_j^T g + g_1(\overline{Z}_{1j}) + g_2(\overline{Z}_{2j}) + ... + g_p(\overline{Z}_{pj}) + a + m_j$, j=1,…,n.

where $g_p(\overline{Z}_{pj})$ is the unknown function for the *p*-th variable of the borrower's risk factors, and a is a constant, which is unable to be identified but is incorporated into the effect of each *Z*. Linton and Nielson (1995), Fan, Hardle, and Mammen (1996) use marginal integration to estimate the univariate quantity for

[9] High-dimension graphs are needed in order to visualize the marginal effect of each variable; this can be very difficult and inconvenient.

each component of the additive semiparametric PLR model in equation (10). See appendix for a brief review of the concept of marginal integration method. An important result from applying marginal integration to the additive PLR model (10) is that the asymptotic distribution of *g's* behaves in the same way as if it were an one-dimensional local nonparametric estimator; thus it avoids the so-called curse of dimensionality that plagues many nonparametric applications.

The additive semiparametric PLR formulation allows for a separate treatment for each component of the nonparametric vector. Furthermore the univariate quantity of each component can be illustrated by a graphical representation. We will use the graphical representation to detect the possible non-linear shape of the borrower's risk factors, including LTV and income.

3. Data

3.1 Construction of the database

Three data sets are used in this paper, the GSE Public Use Database (PUDB), HMDA data, and the 1990 Census. PUDB is the primary data source for the GSEs' single-family mortgage purchase and includes cross-section data from 1993-1996. The single family census tract loan level data of the GSE PUDB contains basic information of the loans that purchased by GSEs, including the location of the property and basic demographic descriptors of the borrower. Aggregating the PUDB loan data at the census tract level, we calculate the number of low-income loan purchases in each tract and compute each tract's share of MSA-total low-income purchases. Loan application data from HMDA provides loan level information about the applicants, loan type, loan amount, location of property, as well as institutional variables. Aggregating the loan data at the census tract level, we can get the number of loan applicants, number of loan applicants eligible for each of the affordable housing goals, numbers of different types of loans, distribution of loan amounts, distribution of borrower incomes, etc.[10]

Using geographical identifiers, we merge the HMDA and GSE PUDB together at the census tract level. Additionally we merge data from the 1990 Census STF3A file to obtain social-economic information for each census tract. Basing on the Office of Management and Budget (OMB)'s definition of MSA and PMSA, we choose census tracts and central cities of the largest 20 MSA/PMSAs. The number of GSE loan purchases in the studied areas was about 900,000 in 1996, which accounts for about one-third of total GSE loan purchases nationwide. On average, 37% of the GSE loans meet the low- and moderate-income goal requirement.

3.2 Visual presentation of the dependent variable

Visual representation of the data is typically more informative when spatial issues are the focus of interest. We present the data in maps using Geographic Information Systems (GIS) technology. The geographical boundaries' data are obtained from the First Street Dataset distributed by Wessex, Inc. The First Street Data is transferred from the geocoded TIGER files of the Bureau of the Census, providing us with the boundary and location of each census tract, place, county, MSA, and State.

Map 1 plots the spatial distribution of the loan purchases as defined by *NPR* in equation (7) for the Baltimore metropolitan area in 1996. The boundary of major central cities is highlighted. Each tract is classified according to the following rules: (a) if the tract's NPR<0.5, so that the tract's share of low income loan purchases is less than half its share of goal-eligible applicants, the tract is termed 'highly under-represented;' (b) if 0.5<NPR<0.8, the tract is labeled 'under-represented;' (c) if 0.8<NPR<1.25, the tract is termed 'closely matched;' (d) if 1.25<NPR<2.0, then the tract's share of loans purchased well exceeds it

[10] HMDA data also contain GSE purchase information; however, it is almost impossible to merge the two data sets at the loan level. Further the coverage of GSE purchases in HMDA dataset is not complete and it is estimated that HMDA contained only 70-75% of the Freddie Mac's purchases (Berkovec and Zorn, 1996).

share of goal-eligible borrowers, and the tract is called 'over-represented;' and (e) if NPR>2, the tract is termed 'highly over-represented;'

Baltimore is a good representative in the sense that it is similar to virtually all other areas in containing the following patterns in the data. The vast majority of tracts in the center cities are at least under-represented, while most of the remote suburban tracts are over-represented. The activity in the center cities is not uniform, and activity in suburban tracts does vary greatly as well. The picture is mixed for the inner ring suburban tracts.[11]

[Add Map 1 here]

[Add Table 1 here]

It is easy to show that the larger the standard deviation of *NPR*, the more extreme the loan purchases mismatch the goal-targeted borrowers spatially. Table 2 reports the means and standard deviations for *NPR* by metropolitan area. These results indicate that the standard deviations of *NPR* tend to have risen slightly over time. Within the same year, there is considerable heterogeneity across metropolitan areas. Boston, San Diego, and Tampa have the least spatial mismatch. A group of areas, including Dallas, Denver, Houston, New York, and Phoenix, have consistently large standard deviations for *NPR* across tracts.

3.3 Explanatory variables

The explanatory variables include a variety of tract traits, loan applicant characteristics, and loan type information. Table 2 lists all of the explanatory variables, their definitions, means and standard deviations. Center city neighborhoods, heavy minority concentration neighborhoods, and poor neighborhoods have been the focus of housing policies, and they have been used to defined the underserved areas in HUD's GSE rulings. If lenders or GSEs view these neighborhoods as riskier after discounting the distribution of borrower's credit risks and try to have the least exposure in these tracts, our model would generate negative coefficients on *C_CITY* and *BLACK* and a positive coefficient on *TR_RATIO*. An interaction of central city with black (*CC_BLCK*) is included to ascertain whether or not if there is any difference in the racial effect on suburban locations as opposed to those in the city.[12]

Because loan level data on risk factors such as LTV, payment-to-income ratio (PTI), or credit score are not available in our data source, we use the mean of goal-eligible applicants' income (*G1IN_LOG*) and the mean LTV (*G1_LTV*) to approximate borrowers' risk distribution. The applicant's LTV is obtained by dividing the loan amount by an estimated house value. The estimated house value is based on the tract median house value and adjusted by the applicant's relative income in the tract median, assuming a constant housing demand elasticity.[13] See Appendix 2 for the details of on the construction of LTV. Since PTI is directly related to borrower's income and credit score is significantly correlated with income (see Pennington-Cross and Nicolas [1999]), *G1IN_LOG* and *G1_LTV* are good approximations of individual risk factors.

Loan type information *G1REFI_RA* and *G1INV_RA* are also included to control the risks associated with different types of loans. Refinance loans typically have a smaller LTV and may have a different risk schedule than original loans. In addition, investor loans are typically considered more risky than owner loans, and GSEs typically purchase very few of them (see DiVenti, 1998). Finally, since the GSEs are not allowed to purchase FHA/VA insured or guaranteed loans, the presence of FHA/VA loans directly implies a smaller share of GSE purchases. We use *G1FHA_RT* to control this and expected a negative effect.

4. Semiparametric regression result

[11] Maps for all other areas show a very similar pattern. Maps for other years (1993-1995) show little inter-temporal variance in terms of the intra-metropolitan area distribution of mortgage purchase activity. See Hu (1999).

[12] In making mortgage purchases, the GSEs are prohibited from considering the location of dwellings, or the age of the neighborhoods where dwellings are located, in a manner that has a discriminatory effect. Analysis of the GSEs' adherence to this prohibition is beyond the scope of this paper, and it would be inappropriate to infer anything regarding GSEs' discrimination from this paper.

[13] Studies show that the income elasticity of housing demand is pretty much invariant, see Mills (1999).

Equation (10) is estimated using census tract-level data from twenty metropolitan areas. Because local market conditions vary and underwriting criteria may not be uniformly enforced across markets, we estimate a fully-interacted model, i.e., a regression for each MSA,[14]

$$(11)\quad NPR_{jc} = b_{1c} *C\_CITY_{jc} + b_{2c} * BLACK_{jc} + b_{3c} *CC\_BLCK_{jc} + b_{4ct} *TR\_RATIO_{jc} + b_{5c} * G1FHA\_RA_{jc}$$

$$+ \boldsymbol{b}_{6c} * G1INV\_RA_{jc} + \boldsymbol{b}_{7c} * G1REFI\_RA_{jc} + g_1(G1IN\_LOG_{\mathrm{jc}}) + g_2(G1\_LTV_{jc}) + \boldsymbol{a}_c + \boldsymbol{e}_{jc}$$

where all variables are as described above and subscription *c* denotes MSA.

4.1.Comparison of PLR with linear and quadratic models

We first compare the partial linear semiparametric model with several other specifications to see if the semiparametric approach can improve the predictive power of the regression. In the PLR model, we use a kernel smoother for the nonparametric part. The Gaussian kernel $k(t) = (2\boldsymbol{p})^{-1/2}\exp(-t^2/2)$ is used and the bandwidths, h_1 for *G1IN_LOG* and h_2 for *G1_LTV*, are chosen by cross-validation (see Appendix 2 for cross-validation function). Table 3 compares the PLR model with a variety of OLS estimations including two linear models and a quadratic model, for Baltimore 1996.

Table 3 here

The first column presents a simple linear model. From this model we can see that *C_CITY* is significantly negative, *BLACK* is significantly negative, and *CC_BLCK* is significant and positive--which offset some of the *BLACK*'s negative effect. *TR_RATIO* has a significant positive effect, indicating that wealthy tracts tend to have higher GSE purchase rates of low-income mortgages. The control variables,

[14] Capozza, Kazarian and Thomson (1997) illustrated the disadvantage of aggregating mortgage default data across metropolitans. Our data also shows that by pooling all the MSAs together, the goodness of fit is significantly lower than the fully interacted model.

GAFHA_RA and *G1INV_RA*, are both negative and significant. *G1REFI_R* is not significant in any of the four models, suggesting that the refinance loan may have the same risk schedule as non-refinance loans. Borrower's income is significant and positive, while LTV is not significant. The second and the third columns of Table 3 present the results of the model augmented by dummies or powers of *G1IN_LOG* and *G1_LTV*; in terms of the statistical significance, the quadratic model doesn't seem to provide any noticeable improvement over the simple linear specification, while the linear dummy model does a little bit better. The positive and significant coefficient on the *SPE_LOW* dummy reveals that the very low-income family may actually have a higher purchase rate. This may be because GSEs are trying to fulfill the special affordable goal.

The last column presents the result of a PLR model with *G1IN_LOG* and *G1_LTV* entering the model nonparametrically. While the coefficients of the other variables are mostly consistent with the linear model, the PLR model's goodness of fit is far superior to the linear model. More importantly, the *C_CITY*'s coefficient becomes insignificant. This implies that, after a better control of applicant's credit risk distribution, center-city location no longer has a significant effect on GSE low-income purchase rates. Because PLR does not impose any specification on applicant's income and LTV, it represents a better control over the non-linearity of these two variables. Comparing the PLR with all the three OLS specifications, we can conclude that PLR does a better job than linear or quadratic specification.[15]

4.2 Neighborhood effects in other metropolitan areas

[15] We also run regressions comparing the GSE lower-income purchases with primary lower-income loan origination. In Appendix 2, we define a dependent variable as $NPR_j^{(o)} \equiv \frac{l_j}{o_j} \Big/ \frac{\sum_{j=1}^{n} l_j}{\sum_{j=1}^{n} o_j}$, where o is the number of lower-income loans originated by primary lenders. Comparing Table A2 and Table 3 and Table A2, we found the patterns are very similar. It suggests that in responding to the neighborhood risks, GSEs and primary lenders have similar behaviors.

We then run the PLR model over each of the 20 metropolitan areas for 1996; the results are presented in Table 4. As a comparison, linear models for each MSA are presented in Table A2 in the Appendix. Compared to the linear model, the goodness of fit of PLR is significantly improved in all the metropolitan areas. Although the two results are similar in terms of the signs of coefficients, the significance levels and the marginal effects are frequently different, especially for *C_CITY* and *BLACK*. In addition, because of the mis-specification problem, the linear model frequently over estimates the significance of the neighborhood characteristics.

We begin our discussion by considering the impact of center cities. The disparity in social and economic trends between central cities and their suburbs has long been discussed and debated. If the GSEs perceive central city tracts as relatively risky, the risk mitigation behavior would imply a negative center city effect on GSE lower income loan purchase rate. Of the 20 metropolitan areas, this is indeed the case in four--Chicago, Cleveland, Philadelphia, and Tampa. In these areas, if a census tract is located in the central city, the purchase rate is lower than for an otherwise observationally equivalent tract in the suburbs. However, in five other cities--Boston, Houston, Minneapolis, New York, and San Diego--the coefficient is significantly positive and, in the remaining 11 metros there are no significant effects. Therefore, except for a few areas, center city status generally does not impose a negative effect on the GSE lower income purchases.

Racial differentials in mortgage lending in general are of widespread public policy interest (e.g., see Munnell, et al. 1996; Ladd, 1998) and racial segregation is a fact of life in most metropolitan areas. Columns 3 and 4 of Table 4 report how a tract's concentration of African-Americans (both in level terms and interacted with central city status) is correlated with our measure of spatial mismatch. The coefficient on *BLACK* is statistically negative in 14 of the 20 metropolitan areas. The interpretation of this is that neighborhoods with higher ratios of African-Americans are more likely to be under-represented, other variables being constant. Of special interest is that the coefficients of the interaction of *BLACK* with center city are mostly positive, and statistically significant only when the coefficient on *BLACK* is significant. Thus, the findings indicate that (a) tracts with higher proportions of African-American households tend to be under-represented in terms of Goal 1 loan purchases in our large metropolitan areas, and (b) a neighborhood's racial component has a greater effect in suburban areas than that in center cities.

TR_RATIO reflects the effect of a neighborhood's relative wealth on GSEs' lower income mortgage purchases. Column 6 shows that in all but one metropolitan area (Oakland) the coefficient on *TR_RATIO* is significant and positive, which implies that there is a larger chance for a lower income borrower to be picked up by the GSEs if she buys a home in a wealthier neighborhood. The marginal effect of *TR_RATIO* is typically large in many metropolitan areas. This suggests that the GSEs are purchasing disproportionately more low-income loans in tracts with higher incomes. The findings for this variable are consistent with an interpretation that the GSEs mitigate risk with respect to low-income loan purchases by targeting loans made in tracts with relatively high incomes in almost all metropolitan areas. The findings do indicate that a goal that targets lower-income families need not always promote loan purchases to low-income neighborhoods.[16]

It is noteworthy to mention that the coefficient on *GAFHA_RA* is mostly significant and negative, which confirms the FHA/VA effect;[17] that the presence of investor loans also has a negative coefficient in most of the cases, but often insignificant. The effect of refinance varies a lot across metropolitan areas, with coefficients of 6 significantly positive and 6 significantly negative. Dropping this variable has no material effect on other variables.[18]

[16] This is consistent with Canner's finding from HMDA data (Canner, 1995).

[17] Of course, some might argue that the level of FHA activity is endogenous in many cities. That is, the lack of GSE activity may be more important causally in explaining FHA's presence than the reverse. This is an issue well beyond the scope of this research.

[18] We also run semi-linear regressions for 1993, 1994, and 1995. In general, there are few changes over time. The biggest one involves the effects of center city status. There are more metropolitans with negative center-city effect in 1993, 1994, and 1995 than there are in 1996. In 1996 HUD changed the GSEs' geographically-targeted goal and narrowed the target areas from the center city down to census tracts with lower income and high minority

It is important to discuss the possible biases due to omitted variables such as credit score, employment history, etc. Because credit score and employment history are important credit risk measurements and they are correlated with neighborhood income and racial component, it is possible that the estimations on *TR_RATIO* and *BLACK* are over-estimated. However, given the strong significant level and large marginal effect of *TR_RATIO*, it is reasonable to expect that the neighborhood income effect still be significant if the omitted variables bias could be fixed. Furthermore, the credit worthiness (measured by credit score and employment history) of suburban applicants should be no lower than that of center city applicants; therefore, our estimation on the *CC_BLCK* is not biased upward. That says our conclusion that a neighborhood's racial component has a greater effect in suburban areas than that in center cities should be affected by the omitted variables bias, if not stronger.

5. Non-linearity of borrower's income and LTV effects.

Having established that the PLR specification is the most preferred formulation, we proceed to estimate the additive PLR model as given by equation (10) with the two non-linear components--applicant's income and LTV. Following Linton and Nielsen (1995), we use the standard normal kernel and the bandwidth is chosen by cross-validation. The estimates of the non-linear components for the logarithms of applicant's income and LTV in Baltimore are presented graphically in Figures 1 and 2. The remainder of the results is presented in the Appendix.

These graphics show the univariate quantities of each variable. The Y-axis is the simulated *NPR* using the additive PLR model, defined as

$$(12) \qquad \overset{\wedge}{NPR}(G1IN\_LOG_j) = E(X_j^T b + s_j^T g) + Eg_2(G1\_LTV_j) + a + g_1(G1IN\_LOG_j) + m_j \text{, or}$$

concentration. Prior to 1996, GSEs could purchase loans from center city rich families to fulfill this goal, and as a result the center city poor families were more likely to be left behind.

$$\hat{NPR}(G1\_LTV_j) = E(X_j^T b + s_j^T g) + Eg_1(G1IN\_LOG_j) + a + g_2(G1\_LTV_j) + m_j \,.$$

From Figure 1 we can see that the relation of applicant's income with GSE purchase rate is not linear but a check-mark "√" shaped curve. When the mean income of applicants is higher than $25600 (with logarithm around 10.15), the higher the income, the greater the GSE purchase rate, which is in agreement with a linear model. However, there is a turning point at about $25600, which is approximately 60% of the area median income of $42206. When an applicant's income is lower than 60% of the AMI, the application will be the target of the special affordable goal (Goal 3). The lower a tract's G1IN_LOG, the more Goal 3-eligible applicants and, therefore, the higher the GSE purchase rate in this tract. This pattern, a higher *NPR* at the lower end of income range, is very consistent across most of the 20 metropolitan areas. See figures in Appendix for more cities.

Figure 2 shows the non-linear relation between tract median LTV and *NPR*. The majority of the LTV lies in the range of 30-100% and the relation between LTV and *NPR* shows a "U-shaped" curve. When the tract median LTV below 80%, the higher the LTV, the lower the purchase rate, which is consistent with our risk-mitigation story. However, when the LTV higher than 80%, the higher LTV, the higher the GSE purchase rate.

This U-shaped relation between LTV and purchase rate appears in 11 out of the 20 metropolitan areas, including Chicago, Minneapolis, San Diego, and Philadelphia. Although the turning points in these cities are not exactly the same, they are all quite close to 70-80%. See the Appendix, Figure A2. This robust result suggests that GSE purchase rate is actually not particularly lower at the high end of LTV.[19]

A possible explanation for this result may rest on the mortgage insurance factor. For a conventional loan with a LTV higher than 80%, private mortgage insurance is typically required. As

Canner and Passmore (1995) point out, the actual decision of whether or not a mortgage is originated rests upon the entity that ultimately bears the credit risk and PMIs bear the majority of the credit risk when a loan is insured. In tracts with a higher mean LTV, there are more loans insured and, therefore, there is a higher chance to be picked up by GSEs. In addition, it is likely that low down-payment borrowers are also low-income borrowers and are being targeted by GSEs to meet their special low-income goal.

That said, we should note that this "U-shaped" pattern does not appear in all of our studied areas. Because the lending criteria are not uniform across metropolitans, it is not surprising to see variations of correlation between LTV and GSE purchase rate. Figure 3 presents three types of the relations. In 3(c) that of Chicago, there is a straight decrease before tract LTV reaches 80%, then it increases slightly; this is essentially the same pattern demonstrated in the Baltimore case. In 3(a), that of Washington, the *NPR* experiences a flat stage before the tract mean LTV reaches 45%, then a decrease. This is consistent with a risk mitigation story. The linear model in Appendix fails to catch the flat part. In 3(b), that of Tampa, there is a flat period before LTV reaches 60% and then it increases; in the linear model, only an increase sign is shown (see Table A1 in Appendix).

6. Conclusion

Using a newly developed additive semiparametric model, this paper investigates how the implicitly-subsided affordable housing credit in the secondary mortgage market is distributed over lower income homebuyers. Particular attention is given to the question of whether (and if so, how) neighborhood characteristics play a role in addition to applicant's credit risk factors and loan types, in determining GSEs lower income mortgage purchases. GSE and HMDA data for the 20 largest MSAs suggest there is a mismatch between the spatial distribution of lower income loan purchases and the lower income loan

[19] While this is consistent with an adverse selection story--that is, because the asymmetric information, GSE was offering loans with higher risks from primary lenders--the results from other variables certainly does not support it.

applicants. Semiparametric regression results show that applicant's risk distribution, loan type frequency, and neighborhood's income and racial status contribute significantly to this mismatch.

The partial-linear (PLR) semiparametric model does a better job, compared with linear and quadratic models, in controlling the non-linear effects of borrower's credit risk factors. The PLR model significantly improves the goodness of fit and it reduce the estimation bias that is found in a linear model. For example, the central city's effect that is found in a linear model generally becomes insignificant in the PLR model.

Detailed PLR analysis is conducted for each of the 20 metropolitan areas. The results suggest that neighborhoods with a higher ratio of African-Americans are more likely to be under-represented and a neighborhood's racial component has a greater effect in suburban areas than that in center cities. Strong evidence is also found that the GSEs purchase disproportionately numbers of lower income loans in relatively affluent neighborhoods. Higher frequency of investor loans and FHA/VA activities also contribute to the spatial mismatch.

The paper investigates the non-linearity of the effects of borrower's risk factors on the GSE lower income purchases, using graphic presentations of the semiparametric results. A check-mark "√" shaped relation between applicant's income and GSE purchase rate is found. When the applicants' income is higher than a certain level, the GSE purchase rate increases with the tract mean of applicants' income. However, there is a higher GSE purchase rate at the lower end of income range. This is likely to be due to the effect of the special affordable goal (Goal 3).

The LTV effect is more complicated. In most metropolitan areas, when the tract median LTV is below 80%, the higher the LTV, the lower the purchase rate, which is consistent with a risk-mitigation story. However, when the LTV is higher than 80%, higher LTVs are typically associated with higher GSE purchase rates. Possible explanations may rest on the introduction of mortgage insurance, which may reduce the risk of high LTV loans borne by the GSEs. Other non-linearity and discontinuity in the mortgage market may further complicate the relation.

While these patterns are broadly consistent across metropolitan areas in terms of the Low and Moderate Income Goal, there is noteworthy variation in behavior across areas. We suspect that detailed examination of this heterogeneity will prove fruitful to understand the spatial affects identified here.

References

Anselin, L. and A. Can. (1998). "Spatial Effect in Models of Mortgage Origination," paper presented in the 43rd North America Meetings of Regional Science Association International.

Archer, W. R., D. H., Gratzlaff and D.C. Ling. (1996). "Measuring the Importance of Location in Housing Price Appreciation," *J. of Urban Economics* 40: 334-353.

Barnett, W. A., J. Powell, and G. Tauchen (eds.). (1991). *Nonparametric and Semiparametric methods in Econometrics and Statistics*, Cambridge University Press: New York.

Berkovec, J. and P. Zorn. (1996). "How Complete is HMDA?: HMDA Coverage of Freddie Mac Purchases," *J. of Real Estate Research*, 11(1), 39-55.

Calem, P.S. and S.M. Wachter. (1999) "Community Reinvestment and Credit Risk: Evidence from An Affordable-Home-Loan Program," *Real Estate Economics*, 27(1), 135-168.

Canner, G. B. (1995). "Home Purchase Lending in Low-Income Neighborhoods and to Low-Income Borrowers," *Federal Reserve Bulletin*, 81(2):71-103.

Capozza, D. R., D. Kazarian, and T.A. Thomson. (1997). "Mortgage Default in Local Markets," *Real Estate Economics*, 25(4), 631-655.

Case, K. E. and R. J. Shille. (1989). "The Efficiency of the Market for Single Family Homes," *AER*, 79(1), 125-137.

Clapp, J. M. and C. Giaccotto. (1994). "The Influence of Economic Variables on Local House Price Dynamics," *J. of Urban Economics* 36(2): 161-183.

Crone, T. M. and R. P. Voith. (1999). "Risk and Return within the Single-Family Housing Market," *Real Estate Economics*, 27(1), 63-78.

DiVenti, T. R. (1998). "The GSEs' Purchases of Single-Family Rental Property Mortgages," U.S. Department of Housing and Urban Development, *PD&R Working Paper* No. HF-004.

Fan, J., W. Hardle, and E. Mammen. (1996). "Direct estimation of Low Dimensional Components in Additive models," mimeo.

Galster, G. (1992). "Research on Discrimination in Housing and Mortgage Markets: Assessment and Future Direction," *Housing Policy Debate*, 3(2), 639-83, 1992.

Gyourko, J. and D. Hu. (1999). "The Intra-Metropolitan Area Distribution of GSE Mortgage Purchases Made in Support of Low-income Related Goals." *Wharton Real Estate Center Working Paper* No. 317.

Hardle, W. (1990). *Applied Nonparametric Regression*, Cambridge University Press: New York.

Harrison, D. (1999). "The Importance of Lender Heterogeneity in Mortgage Lending," mimeo.

Hu, D. (1999). "Mapping the Intra-Metropolitan Distribution of the GSE Purchases Made in Support of the Affordable Housing Goals," *Wharton Real Estate Center Working Paper*, No. 317 Supplement.

Kain, J. F. (1968). "Housing Segregation, Negro Employment, and Metropolitan Decentralization," *Quarterly Journal of Economics*, May, 82, 175-97.

LaCour-Little, M. and R. Green. (1997). "Are Minorities or Minority Neighborhoods More Likely to Get Low Appraisals?" *J. of Real Estate Finance and Economics*, 16(3).

Ladd, H. (1998). "Evidence on Discrimination in Mortgage Lending." *The Journal of Economic Perspectives*, 12 (2), 41-62.

Lang, W.W. and L. Nakamura. (1993). "A model of Redlining," *J. of Urban Economics* 33: 223-234.

Linton, O. B. and J. P. Nielsen. (1995). "A Kernel Method of Estimating Structural Nonparametric regression based on Marginal Integration," *Biometrika*, 82, 93-100.

Liu, Z. and T. Stengos. (1999). "Non-linearities in Cross-Country Growth Regreesions: A semiparametric approach," *Journal of Applied Econometrics*, 14, 527-538.

MacDonald, H. (1996). "Expanding Access to the Secondary Mortgage Markets: The Role of Central City Lending Goals." *Growth & Change*. 27(3), 298-312.

Manchester, P. B. (1998). "Characteristics of Mortgages Purchased by Fannie Mae and Freddie Mac, 1996-97 Update". U.S. Department of Housing and Urban Development, *PD&R Working Paper* No. HF-006.

Mills, E.S. and L.S. Lubuele (1996). "Performance of Residential Mortgage Loans in Low and Moderate Income Neighborhoods," *Journal of Real Estate Finance and Economics* 9: 245-262.

Munnell, A. H., G. M. Tootell, L. E. Browne, and J. McEneaney. (1996). "Mortgage Lending in Boston: Interpreting HMDA Data," *American Economic Review*, 86 (1), 25-53.

Pennington-Cross, A. and J. Nichols. (1999). "Credit history and the FHA-Conventional Choice." *Wharton Real Estate Working Paper*, No. 319.

Robinson, P. (1988). "Root-N-Consistent semiparametric regression," *Econometrica*, 56, 931-935.

U.S. Department of Housing and Urban Development. (1995). "The Secretary of HUD's Regulation of the Federal National Mortgage Association and the Federal Home Loan Mortgage Corporation, Final Rule," *Federal Register*, 60, 61846-62005.

Van Order, R. (1996). " Discrimination and the Secondary Market." In John Goering and Ron Wienk, eds. *Mortgage Lending, Racial Discrimination, and Federal Policy*. Washington, D.C.: Urban Institute Press, 143-64.

Van Order, R., A.-M. Westin and P. Zorn. (1993). "Effects of the Racial Composition of Neighborhoods on Default, and Implications for Racial Discrimination in Mortgage Markets," Mimeo, Freddie Mac.

Appendix 1. Additive semiparametric model

Considering a simple version of Equation (9),

(A1) $Y_j \equiv NPR_j = X_j^T \boldsymbol{b} + g(Z_j) + \boldsymbol{e}_j$.

where X is a vector of neighborhood characteristics, and Z is a vector of borrower's credit risk factors. Because Z and X are correlated, estimators of $\boldsymbol{b}$ based on incorrect parameterization of g are generally inconsistent and misleading. Robinson (1988) provided a partial linear semiparametric method to get a $\sqrt{N}$ -consistent estimation of $\boldsymbol{b}$. The first step of his method is to use the nonparametric kernel method (see Hardle [1990] for an comprehensive and accessible description of kernel smoothing techniques) to get conditional expectations of $\hat{Y}_j = E(Y_j \; | Z_j)$ and $\hat{X}_j = E(X_j | \; Z_j)$. We denote $\tilde{Y} \equiv Y - \hat{Y}$ and $\tilde{X} \equiv X - \hat{X}$. The second step is to run OLS of $\tilde{Y}$ on $\tilde{X}$ to estimate $\boldsymbol{b}$,

(A2) $\hat{\boldsymbol{b}} = (\tilde{X}^T \tilde{X})^{-1} \tilde{X}^T \tilde{Y}$

Robinson proved that the $\hat{\boldsymbol{b}}$ is $\sqrt{N}$ -consistent, which means the $\hat{\boldsymbol{b}}$ will asymptotically converge to its true value at the order of $\sqrt{N}$, where N is the number of observations.

Now suppose Z=(z_1, z_2) , where z_1 and z_2 are two borrower's risk factors, e.g., income and LTV. Further we denote $g(z_{1j}, z_{2j}) \equiv Y_j - X_j^T \hat{\boldsymbol{b}}$, and impose that

(A3) $g(z_{1j}, z_{2j}) = g_1(z_{1j}) + g_2(z_{2j}) + \boldsymbol{a} + u_j$.

Linton and Nielson (1996) provided a simple kernel method based on marginal integration that estimates the relevant univariate quantity in an additive nonparametric regression. Fan, Hardle, and Mammen (1996) extended the regression function of A3 to allow for a more general partial linear formulation.

The idea behind marginal integration is that given the joint estimation $\hat{g}(z_{1i}, z_{2j})$ and the additive assumption of $g_1(z_1)$ and $g_2(z_2)$, one can obtain an estimator of $g_1(z_1)$ plus a constant by integrating $\hat{g}(z_{1i}, z_{2j})$ over z_2. Instead of estimate $g_1(z_1)$ we estimate $g_1(z_1)$ plus a constant, *i.e.*,

(A4) $Q_k(z_k) \equiv g_k(z_k) + c,$ k=1,2

where $c = E[g_2(x_2) + \alpha] = E[\,g_1(x_1) + \alpha]$.

Let $\hat{G}$ be the $n \times n$ matrix of estimates, with typical element $\hat{g}(z_{1i}, z_{2j})$ calculated using the Gaussian kernel $k(t) = (2\pi)^{-1/2}\exp(-t^2/2)$ and bandwidth h_1 and h_2. Define the $n \times 1$ vectors $\hat{Q}_1,\ \hat{Q}_2$ containing the estimated univariate components evaluated at each sample point

(A5) $\hat{Q}_1 = \hat{G}\,q_1,\ \hat{Q}_2 = \hat{G}^T\ q_2.$

where q's are the empirical distribution functions for weighting, that is $q_{1,}\ q_2 = (1,1,...,1)^T / n$. Estimations from (A5) are utilized in Equation (12) in the simulation. A recent application of additive partial linear model (Liu & Stengos, 1999) illustrated the effectiveness of the marginal integration method in revealing the non-linearity among variables.

Appendix 2. Cross-validation

The problem of deciding how much to smooth is of great importance in nonparametric and semiparametric regression. In the kernel smoothing case that is used in this paper, the accuracy depends mainly on the selection of bandwidths. Several bandwidth selection procedures were discussed in Hardle (1990). The basic idea behind all the bandwidth selection procedures are to minimize a measurement of the estimation bias (deviations of the conditional expectation from the observed value) and variances. Cross-validation function is one of these measurements, which is defined as

$$\text{(A6)} \qquad CV(h) = n^{-1}\sum_{j=1}^{n}[Y_j - \hat{m}_{h,j}(X_j)]^2\, w(X_j)$$

where $\hat{m}_{h,j}(X_j)$ is a kernel estimation of Y in which one, say the j-th observation is left out, and $w(X_j)$ is a non-negative weight function which we use the empirical weighting matrix $(1,1,...,1)^T/n$. The function validates the ability to predict $\{Y_j\}_{j=1}^{n}$ across the sub-samples $\{X_{i,}Y_i\}_{i\neq j}$. Figure A3 plots the cross-validation function vs the bandwidths for the data of Baltimore. We choose our bandwidth at the lowest point of CV.

Appendix 3. Construction of LTV

In HMDA data we have the loan amount but not the house value for each loan application. To estimate the LTV, we need to estimate a house value for each loan application. A reasonable way is to use the applicant’s income to estimate his house value. According to Mills (1999), the income elasticity of housing demand (house value) is pretty much invariant and the elasticity is less than 1. We assume the income elasticity of house value within one census tract is constant, e_j, and we have the following relations

(A7) $$\frac{HV_{ij} - MedHV_j}{MedHV_j} = e_j \frac{Incm_{ij} - MedIncm_j}{MedIncm_j}.$$

where HV_{ij} is the house value for each applicant that needs to be estimated, $Incm_{ij}$ is the applicant's income, $MedHV_j$ and $MedIncm_j$ are the tract median house value and median income, respectively. Because house value and location are closely correlated, the elasticity within one census tract should be smaller than the elasticity in general. In this paper, we use e_j=0.6.

To eliminate the inflation and house price depreciation/appreciation issue, we adjust the median income here by the Consumer Price Index (CPI), and median house value by the Freddie Mac Repeat Sale Housing Index. Once we get the estimation of each loan application's house value, we can compute the LTV for each loan application and the tract mean of lower income applicants' LTV for each tract. See Table 2 for the distribution of the tract mean LTV.

Table 1: The GSEs lower income home mortgage normalized purchase rate (NPR) of census tracts within each metropolitan area, 1993 and 1996

	1996		1993	
Metropolitan Areas	Mean	Std	Mean	Std
Atlanta	1.10	0.90	0.88	0.55
Baltimore	1.16	1.01	0.92	0.71
Boston	1.09	0.62	0.92	1.00
Chicago	1.22	1.06	0.83	0.67
Cleveland	1.31	1.26	0.90	0.72
Dallas	1.31	1.34	1.03	0.96
Denver	1.21	0.80	0.96	0.63
Detroit	1.34	1.36	0.89	0.83
Houston	1.23	1.28	0.90	0.92
Los Angeles	1.25	0.89	1.02	0.54
Miami	1.04	0.83	0.88	0.57
Minneapolis	1.09	0.68	0.92	0.48
New York	1.30	2.00	0.98	1.56
Oakland	1.18	1.27	0.97	0.49
Philadelphia	1.31	1.00	0.97	0.72
Phoenix	1.13	1.22	0.96	0.71
Pittsburgh	1.29	1.93	1.08	1.58
San Diego	1.11	0.60	0.97	0.41
Tampa	1.15	0.82	0.94	0.61
Washington	1.24	0.97	0.95	0.52

Note: Data are for the 20 largest MSA or PMSA.

Table2. Explanatory variables

Variable	Definition	Mean	Std
Neighborhood traits			
C_CITY	Center city dummy, 1 if a tract locates in a census designed center city, 0 otherwise;	0.484	0.500
BLACK	Ratio of African American households to total number of households in a tract;	0.196	0.307
CC_BLCK	Interaction of *C_CITY* and *BLACK*;	0.149	0.297
TR_RATIO	Ration of tract family median income to the metropolitan area median;	1.019	0.487
Loan type Information			
GAFHA_RA	Frequency of FHA/VA loans in all the lower income loan application in a tract	0.144	0.132
G1INV_RA	Frequency of investor loans (non-owner occupied) in all the lower income loan application in a tract	0.054	0.081
G1REFI_RA	Frequency of refinance loans in all the lower income loan applications in a tract;	0.384	0.196
Borrower's risk distribution			
G1_LOGINC	mean income of lower income loan applications, in log form	10.251	0.175
SPE_LOW	Income dummy, 1 if mean income of loan applicants meets the income criteria for the special affordable goal.	0.270	0.444
G1_LTV	mean LTV of lower income loan applications in a tract	0.657	0.320
LTV80	LTV dummy, 1 if G1_LTV higher than 80%, 0 otherwise	0.197	0.398

Note: Data are for the 20 metropolitan areas, 1996.

Table 3. Comparasion of varies OLS specifications with PLR results, Baltimore 1996 (n=559)

	OLS Linear	OLS with Dummies	OLS Quadratic	PLR
INTERCEP	-18.674	-20.809	520.415	
	(-4.778)	(-5.412)	(2.894)	
C_CITY	-0.231	-0.214	-0.196	-0.16
	(-2.258)	(-2.12)	(-1.915)	(-1.585)
BLCKRATE	-1.370	-1.362	-1.363	-1.37
	(-5.523)	(-5.604)	(-5.54)	(-5.780)
CC_BLCK	1.131	1.052	1.037	1.00
	(4.026)	(3.826)	(3.703)	(3.720)
TR_RATIO	0.965	0.528	0.931	0.86
	(10.6)	(4.333)	(10.242)	(9.431)
GAFHA_RA	-1.747	-1.637	-1.597	-1.87
	(-6.312)	(-6.031)	(-5.678)	(-6.572)
G1REFI_R	-0.153	-0.142	-0.006	-0.33
	(-0.555)	(-0.526)	(-0.021)	(-1.176)
G1INV_RA	-1.238	-1.427	-1.647	-1.62
	(-1.755)	(-2.067)	(-2.323)	(-2.111)
G1IN_LOG	1.905	2.137	-103.157	
	(4.989)	(5.674)	(-2.9401)	
Special Low-inc Dummy		0.548		
		(5.26)		
[G1IN_LOG]^2			5.121	
			(2.995)	
G1_LTV	0.018	0.108	-0.706	
	(0.173)	(0.72)	(-1.692)	
LTV80 Dummy		-0.038		
		(-0.429)		
[G1_LTV]^2			0.363	
			(1.765)	
R-sqr	0.546	0.568	0.554	0.595
Adj. R-sqr	0.539	0.560	0.546	0.589
F-value	73.450	65.420	61.760	101.003

Note: The t-statistics are given in parentheses. The bandwidth was chosen using Cross-validation method. The Special Low-inc Dummy in Model 2 is a dummy for the low-income family who meets the income requirement of the special affordable goal, and the LTV80 Dummy is 1 for all tracts with a 80% or higher LTV.

Table 4: PLR Semiparametric regression result

MSANM	*C_CITY*	BLCKRATE	CC_BLCK	*TR_RATIO*	*GAFHA_RA*	G1REFI_R	G1INV_RA	_RSQ_	N
atl	0.09	-0.66 ***	0.03	0.45 ***	-1.63 ***	0.10	-0.26	0.659	466
	0.91	-4.50	0.15	5.51	-5.57	0.37	-0.51		
bal	-0.16	-1.37 ***	1.00 ***	0.86 ***	-1.87 ***	-0.33	-1.62 **	0.595	559
	-1.59	-5.78	3.72	9.43	-6.57	-1.18	-2.11		
bos	0.14 **	-0.99	0.46	0.16 **	-1.48 ***	0.62 ***	-0.13	0.461	613
	2.38	-1.22	0.57	2.24	-4.35	3.80	-0.34		
chi	-0.23 ***	-1.24 ***	0.45 **	0.32 ***	-1.87 ***	-0.02	-1.32 ***	0.489	1454
	-4.04	-6.79	2.44	5.12	-8.49	-0.12	-3.61		
cle	-0.96 ***	-1.37 ***	1.08 ***	0.35 ***	-2.41 ***	-0.49	-1.08 *	0.546	597
	-7.04	-6.91	4.48	2.98	-3.82	-1.60	-1.68		
dal	-0.13	-1.29 **	1.07 *	0.97 ***	-1.29 ***	0.08	0.22	0.517	527
	-1.10	-2.06	1.68	9.21	-3.77	0.15	0.26		
den	0.06	-1.55 **	1.17 *	1.05 ***	-1.58 ***	-0.84 ***	2.05 **	0.614	406
	0.90	-2.47	1.81	12.81	-6.04	-3.51	2.28		
det	-0.18	-1.33 ***	0.58 **	1.07 ***	-1.19 ***	0.19	-0.65	0.509	1179
	-1.47	-5.68	2.20	10.93	-3.44	0.75	-1.09		
hou	0.38 ***	0.14	-0.68	0.74 ***	-0.19	1.78 ***	-0.56	0.576	666
	4.28	0.30	-1.44	8.73	-0.42	3.34	-0.93		
la	0.00	-0.67 ***	0.02	0.23 ***	-1.68 ***	-0.35 ***	-1.17 ***	0.354	1607
	-0.09	-4.68	0.13	5.67	-10.47	-2.63	-4.88		
mia	0.04	-1.26 ***	0.12	0.31 ***	1.57 ***	-0.87 ***	1.05	0.478	262
	0.34	-6.38	0.48	3.63	3.55	-3.03	1.16		
min	0.20 ***	1.89	-2.29	0.91 ***	-1.23 ***	0.85 ***	-1.10 *	0.561	625
	3.74	1.14	-1.37	11.71	-5.57	3.93	-1.69		
ny	1.01 ***	-0.09	-0.84	0.77 ***	-0.69 **	-0.73 ***	0.15	0.177	2217
	6.46	-0.15	-1.49	9.24	-2.02	-4.94	0.51		
oak	-0.31	-1.32 **	0.58	-0.04	-1.74 **	1.30 ***	-1.18	0.236	453
	-1.29	-2.42	0.90	-0.20	-2.57	2.85	-1.50		
phl	-0.18 **	-1.30 ***	0.70 ***	0.55 ***	-1.88 ***	-0.35 *	-0.13	0.451	1209
	-2.50	-7.55	3.58	7.49	-7.19	-1.90	-0.32		
pho	-0.01	-3.39	2.33	0.46 ***	-2.03 ***	0.57 **	-0.05	0.812	461
	-0.15	-1.58	1.07	6.13	-8.85	2.12	-0.08		
pit	0.02	-0.63	0.29	1.33 ***	-1.27	-1.22 **	-0.76	0.330	628
	0.10	-0.91	0.39	7.41	-1.55	-2.25	-0.65		
sd	0.11 *	-1.86 **	0.87	0.28 ***	-1.02 ***	0.56 ***	-1.00 ***	0.427	429
	1.95	-2.50	1.12	4.13	-4.06	2.81	-2.68		
tam	-0.31 ***	-1.02 ***	0.79 **	0.97 ***	-0.88 **	0.17	1.49 *	0.510	405
	-3.08	-4.82	2.50	8.45	-2.32	0.53	1.67		
was	-0.13	-1.19 ***	0.51 ***	0.43 ***	-1.42 ***	-0.28	-0.16	0.542	888
	-1.49	-9.64	3.09	5.50	-7.30	-1.48	-0.34		

Note: t-statistics is reported under the coefficient. ***significant at 1% level, ** at 5% level, and * 10 % level.
Nonparametric variables are the G1IN_LOG and G1_LTV, and the bandwidths are h=1.28*std(x)*n^(-1/6) for both variables.

Table A1: Linear Regression Result for 20 MSAs 1996

MSANM	INTERCEP	*C_CITY*	BLCKRATE	CC_BLCK	*TR_RATIO*	*GAFHA_RA*	G1REFI_R	G1INV_RA	G1IN_LOG	G1_LTV5	_RSQ_	N
Atl	-35.57	0.08	-0.73 ***	0.23	0.57 ***	-1.51 ***	0.14	0.23	3.55 ***	0.05	0.59	466
	-8.81	0.81	-4.84	1.20	7.07	-5.09	0.51	0.48	9.02	0.50		
Bal	-18.67	-0.23 **	-1.37 ***	1.13 ***	0.96 ***	-1.75 ***	-0.15	-1.24 *	1.91 ***	0.02	0.55	559
	-4.78	-2.26	-5.52	4.03	10.60	-6.31	-0.56	-1.75	4.99	0.17		
Bos	-24.57	0.13 **	-1.15	0.88	0.16 **	-1.99 ***	0.20	-0.48	2.45 ***	-0.09	0.25	613
	-7.48	2.08	-1.22	0.93	2.07	-5.41	1.13	-1.14	7.78	-0.75		
Chi	-12.75	-0.28 ***	-1.35 ***	0.62 ***	0.30 ***	-2.04 ***	-0.15	-1.32 ***	1.42 ***	-0.26 ***	0.45	1454
	-5.55	-5.05	-7.46	3.34	4.89	-9.90	-1.02	-3.64	6.40	-3.65		
Cle	-19.70	-0.95 ***	-1.39 ***	1.20 ***	0.38 ***	-2.65 ***	-0.42	-0.93	2.13 ***	-0.04	0.49	597
	-4.65	-7.46	-6.89	4.97	3.53	-4.39	-1.40	-1.59	5.08	-0.22		
Dal	-36.89	-0.13	-0.84	1.10 *	1.11 ***	-1.31 ***	-0.42	1.20	3.66 ***	0.16	0.44	527
	-6.85	-1.12	-1.31	1.70	10.31	-3.78	-0.84	1.42	6.90	1.34		
Den	-25.12	0.07	-1.29 *	1.02	1.01 ***	-1.33 ***	-0.45 *	2.10 **	2.50 ***	0.08	0.51	406
	-5.36	0.93	-1.89	1.45	11.42	-4.82	-1.79	2.14	5.44	0.81		
Det	-7.21	-0.27 **	-1.37 ***	0.77 ***	1.01 ***	-1.24 ***	0.62 **	-1.37 **	0.76 **	0.03	0.44	1179
	-2.04	-2.18	-5.63	2.82	10.24	-3.86	2.42	-2.28	2.17	0.21		
Hou	-48.61	0.39 ***	-0.03	-0.39	0.71 ***	0.67	2.07 ***	-0.60	4.82 ***	-0.21 **	0.46	666
	-10.77	4.11	-0.06	-0.76	8.06	1.44	3.55	-0.91	10.75	-2.02		
La	-7.57	0.04	-0.75 ***	-0.08	0.25 ***	-1.69 ***	-0.23 *	-1.09 ***	0.90 ***	-0.17 ***	0.23	1607
	-3.63	0.85	-4.91	-0.40	6.00	-10.48	-1.73	-4.60	4.46	-2.76		
Mia	-4.06	0.03	-1.21 ***	0.16	0.28 ***	1.93 ***	-0.71 **	1.36	0.51	-0.07	0.39	262
	-0.64	0.21	-5.95	0.60	3.21	4.31	-2.47	1.58	0.81	-0.51		
min	-23.18	0.17 ***	0.69	-1.01	0.89 ***	-1.11 ***	0.75 ***	0.58	2.26 ***	0.02	0.48	625
	-7.03	3.07	0.39	-0.57	11.31	-5.07	3.52	1.09	6.99	0.21		
ny	-5.93	1.17 ***	-0.05	-0.98 *	0.75 ***	-0.83 **	-0.88 ***	0.15	0.59 ***	0.03	0.13	2217
	-2.98	7.39	-0.09	-1.71	9.13	-2.48	-6.16	0.52	3.05	0.22		
oak	-17.80	-0.38	-1.14 **	0.68	-0.01	-1.86 ***	1.36 ***	-1.01	1.81 **	-0.10	0.16	453
	-2.36	-1.61	-2.25	1.12	-0.08	-2.77	3.07	-1.29	2.52	-0.55		
phl	-11.61	-0.25 ***	-1.37 ***	0.83 ***	0.58 ***	-2.07 ***	-0.59 ***	0.26	1.27 ***	0.09	0.42	1209
	-5.08	-3.41	-8.16	4.42	8.14	-8.32	-3.32	0.78	5.64	1.08		
pho	-43.21	-0.19	-7.75 **	7.75 *	0.37 ***	-2.77 ***	-0.64	-0.34	4.44 ***	-0.02	0.32	461
	-6.08	-1.33	-1.97	1.95	2.67	-7.00	-1.34	-0.29	6.30	-0.16		
pit	-6.49	0.01	-1.27 *	0.77	1.21 ***	-0.97	-0.49	-1.45	0.71	-0.13	0.15	628
	-1.19	0.05	-1.71	0.95	7.17	-1.10	-0.94	-1.14	1.30	-0.60		
sd	-12.81	0.11 *	-1.94 **	0.87	0.32 ***	-0.93 ***	0.69 ***	-0.62 *	1.31 ***	0.07	0.34	429
	-3.98	1.74	-2.52	1.07	4.85	-3.79	3.62	-1.80	4.16	0.84		
tam	-19.16	-0.31 ***	-0.71 ***	0.91 ***	1.04 ***	-0.93 **	-0.50	1.73 *	1.94 ***	0.30 ***	0.43	405
	-3.27	-2.98	-3.42	2.87	8.96	-2.41	-1.56	1.83	3.29	2.61		
was	-22.97	-0.16 *	-1.26 ***	0.72 ***	0.49 ***	-1.37 ***	-0.15	-0.17	2.34 ***	-0.26 ***	0.49	888
	-7.01	-1.76	-10.07	4.51	6.19	-7.21	-0.83	-0.37	7.52	-3.51		

Note: t-statistics is reported under the coefficient. ***significant at 1% level, ** at 5% level, and * 10 % level.

Table A2. OLS and PLR results for comparing the GSE lower-income purchases and primary lenders lower-income origination. Baltimore 1996 (n=556)

	OLS with Dummies	PLR
INTERCEP	-11.13	
	(-2.90)	
C_CITY	-0.21	-0.12
	(-1.884)	(-1.063)
BLCKRATE	-0.91	-0.99
	(-3.427)	(-3.826)
CC_BLCK	0.70	0.66
	(2.288)	(2.237)
TR_RATIO	0.98	0.89
	(9.506)	(8.815)
G1OR_FHA	-1.57	-1.42
	(-6.492)	(-5.687)
G1OR_REF	-0.27	-0.12
	(-0.986)	(-0.418)
G1OR_INV	-0.47	-0.30
	(-0.758)	(-0.471)
G1OR_INL	1.17	
	(3.144)	
G1ORLTV5	-0.09	
	(-0.587)	
SPELOW	0.07	
	(0.656)	
G1ORLTV8	0.07	
	(0.767)	
RSQ	0.45	0.534
F-value	38.3	77.5

Note: The t-statistics are given in parentheses. The bandwidth was chosen using Cross-validation method.
The dependent variable is defined by the number of lower-income loans originated by primary lenders in each tract,

i.e. $NPR_j^{(o)} \equiv \frac{l_j}{o_j} \Bigg/ \frac{\sum_{j=1}^{n} l_j}{\sum_{j=1}^{n} o_j}$, where o is the number of lower-income loans originated by primary lenders.

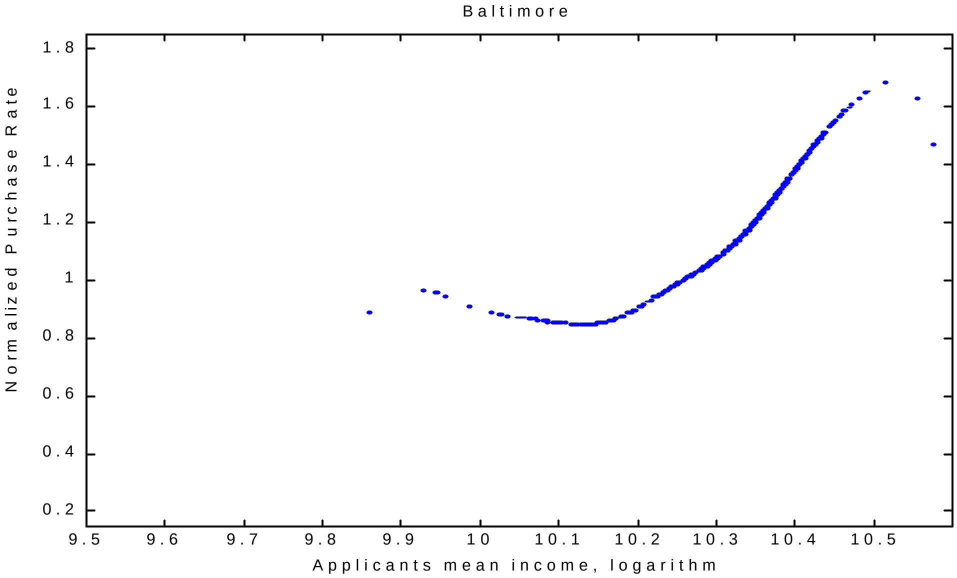


Figure 1. The non-linear relation of lower income applicants' mean income (G1IN_LOG) with normalized GSE purchase rate (NPR), Baltimore 1996. A product of Gaussian kernel is used, and the bandwidths are chosen using Cross-validation method. The *NPR* is simulated using the additive PLR model as in equation 12.

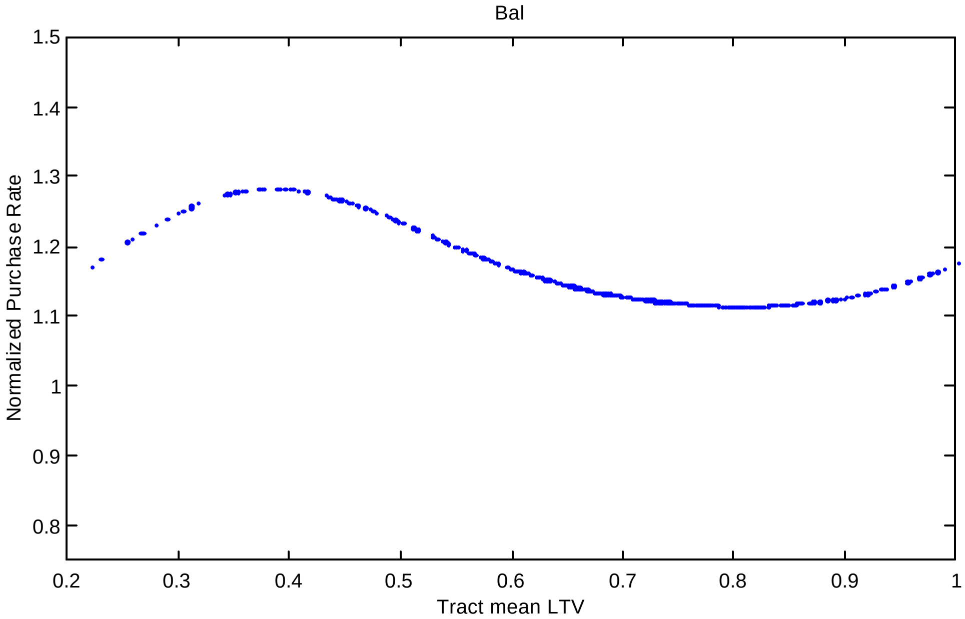
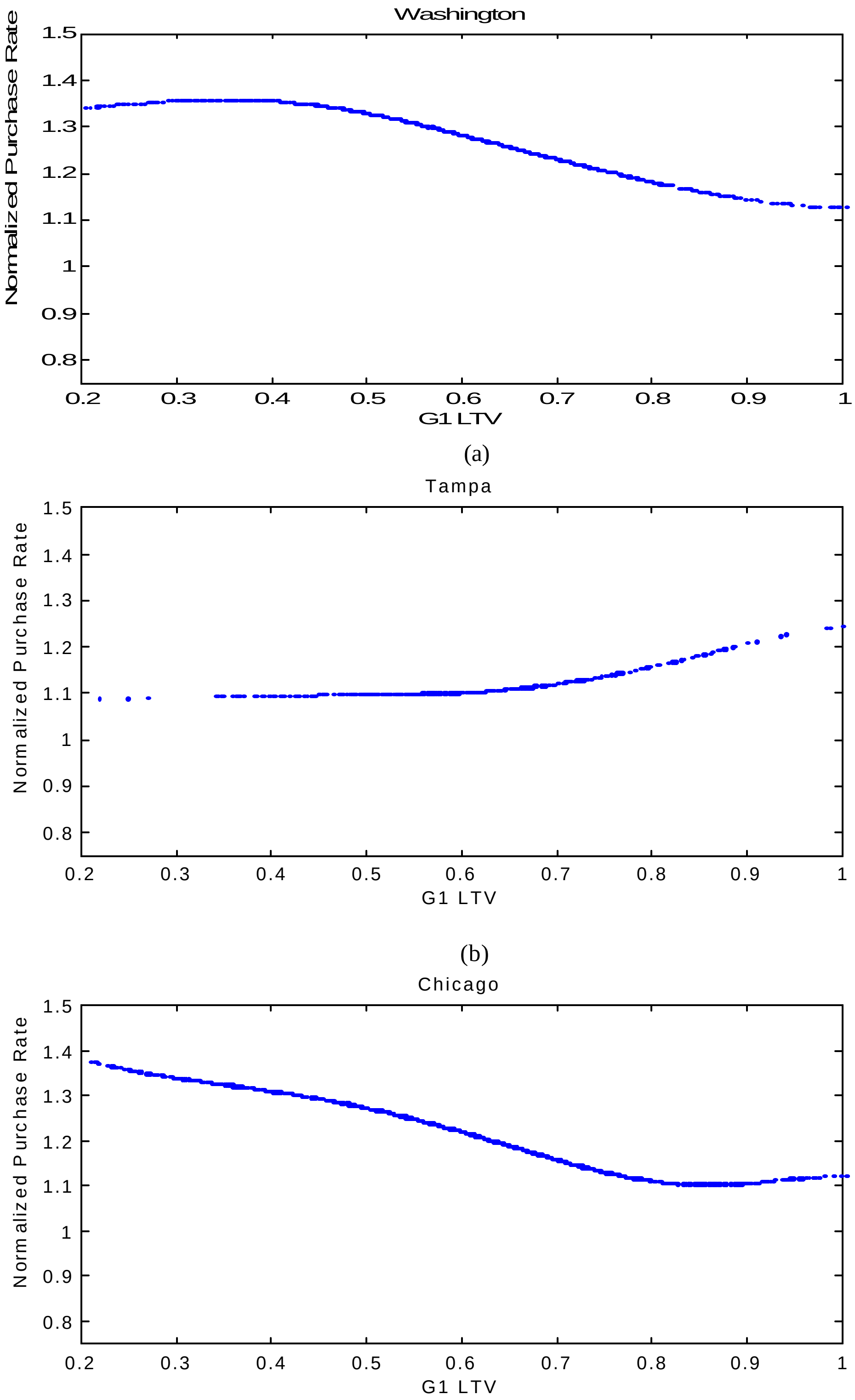


Figure 2. The non-linear relation of lower income applicants' tract mean LTV with normalized GSE purchase rate (NPR), Baltimore 1996. A product of Gaussian kernel is used and the bandwidths are chosen using Cross-validation method. The *NPR* is simulated using the additive PLR model as in equation 12.


Washington
Normalized Purchase Rate
1.5
1.4
1.3
1.2
1.1
1
0.9
0.8
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
G1 LTV
Tampa
Normalized Purchase Rate
G1 LTV
Chicago
Normalized Purchase Rate
G1 LTV

(a)

(b)

(c)

Figure 3. There types of non-linear relation of tract mean LTV (G1_LTV) with normalized GSE purchase rate (NPR). A product of Gaussian kernel is used and the bandwidths are chosen using Cross-validation method. The *NPR* is simulated using the additive PLR model as in equation 12.

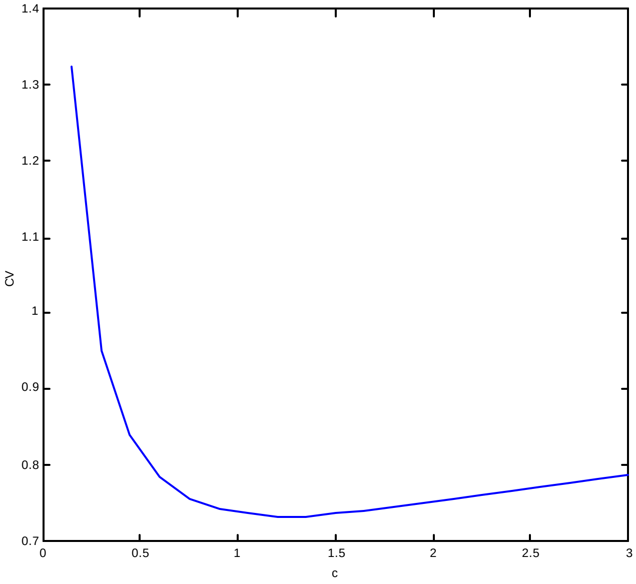


Figure A1. Cross-Validation Function with bandwidth h=c*std(x)*n^(-1/6).

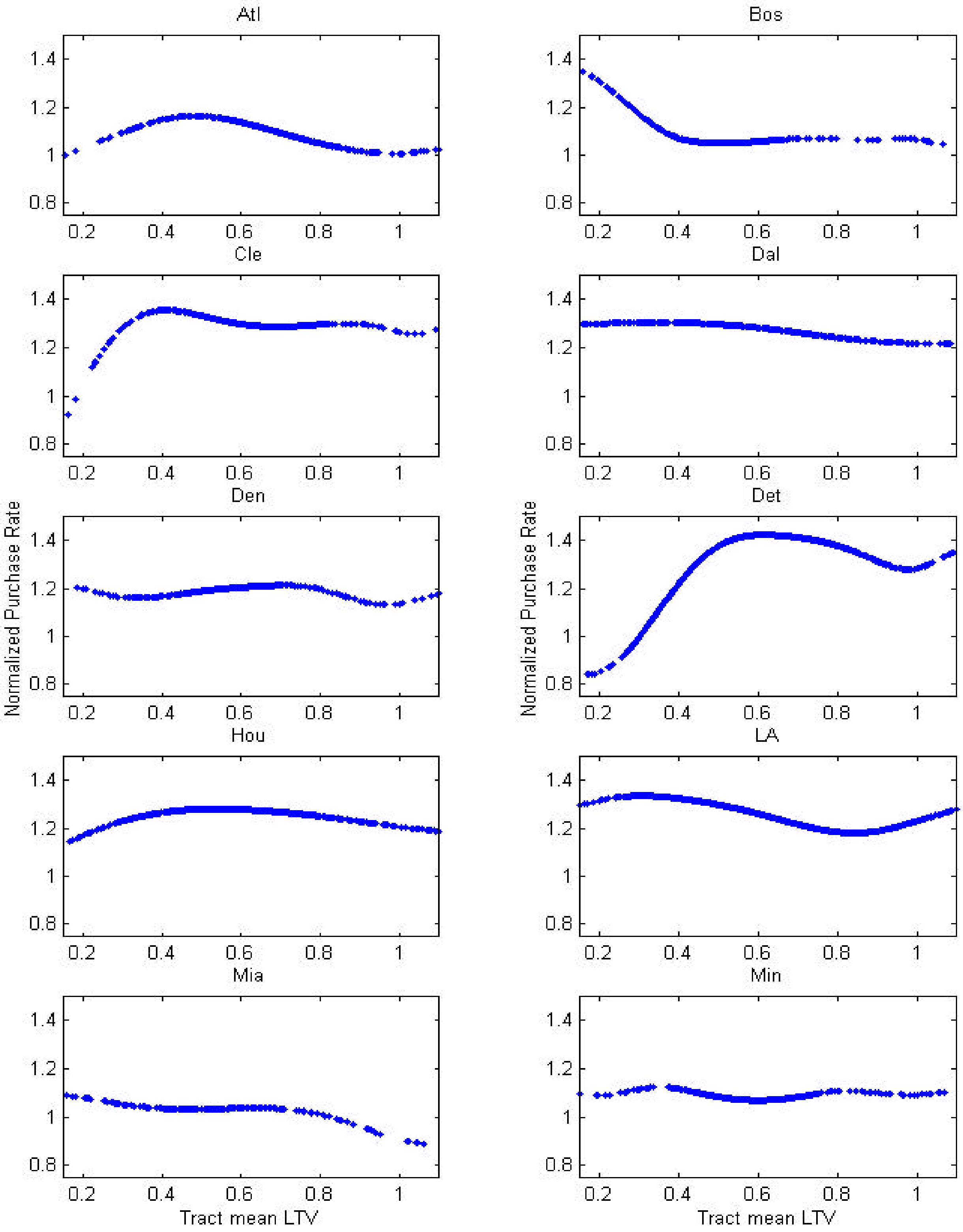

Atl
Bos
Cle
Dal
Den
Det
Hou
LA
Mia
Min
1.4
1.2
1
0.8
0.2
0.4
0.6
0.8
1
Normalized Purchase Rate
Normalized Purchase Rate
Tract mean LTV
Tract mean LTV

Figure A2. The non-linear relation of lower income applicants' tract mean LTV with estimated GSE purchase rate (NPR). A product of Gaussian kernel is used and the bandwidths are chosen using Cross-validation method. The *NPR* is simulated using the additive PLR model as in equation 12.

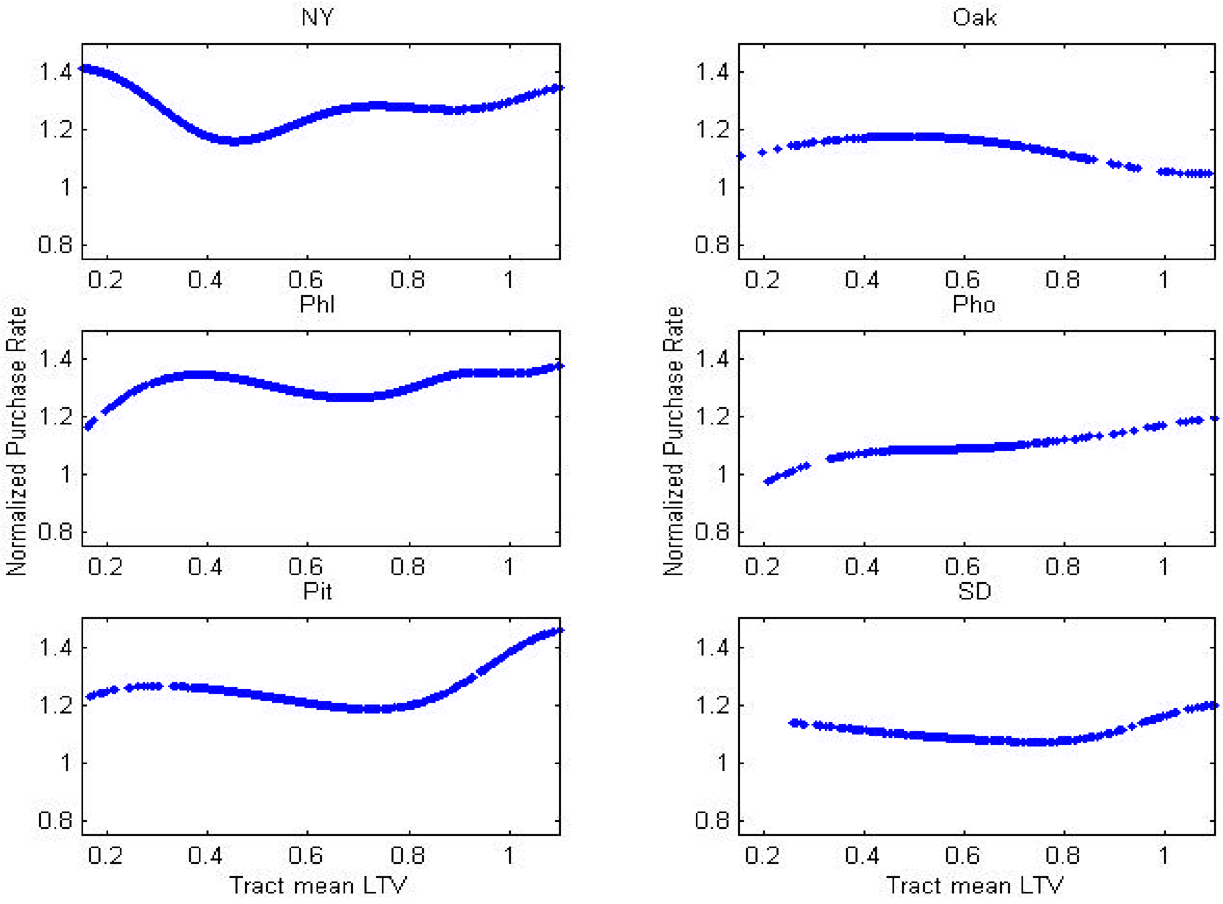
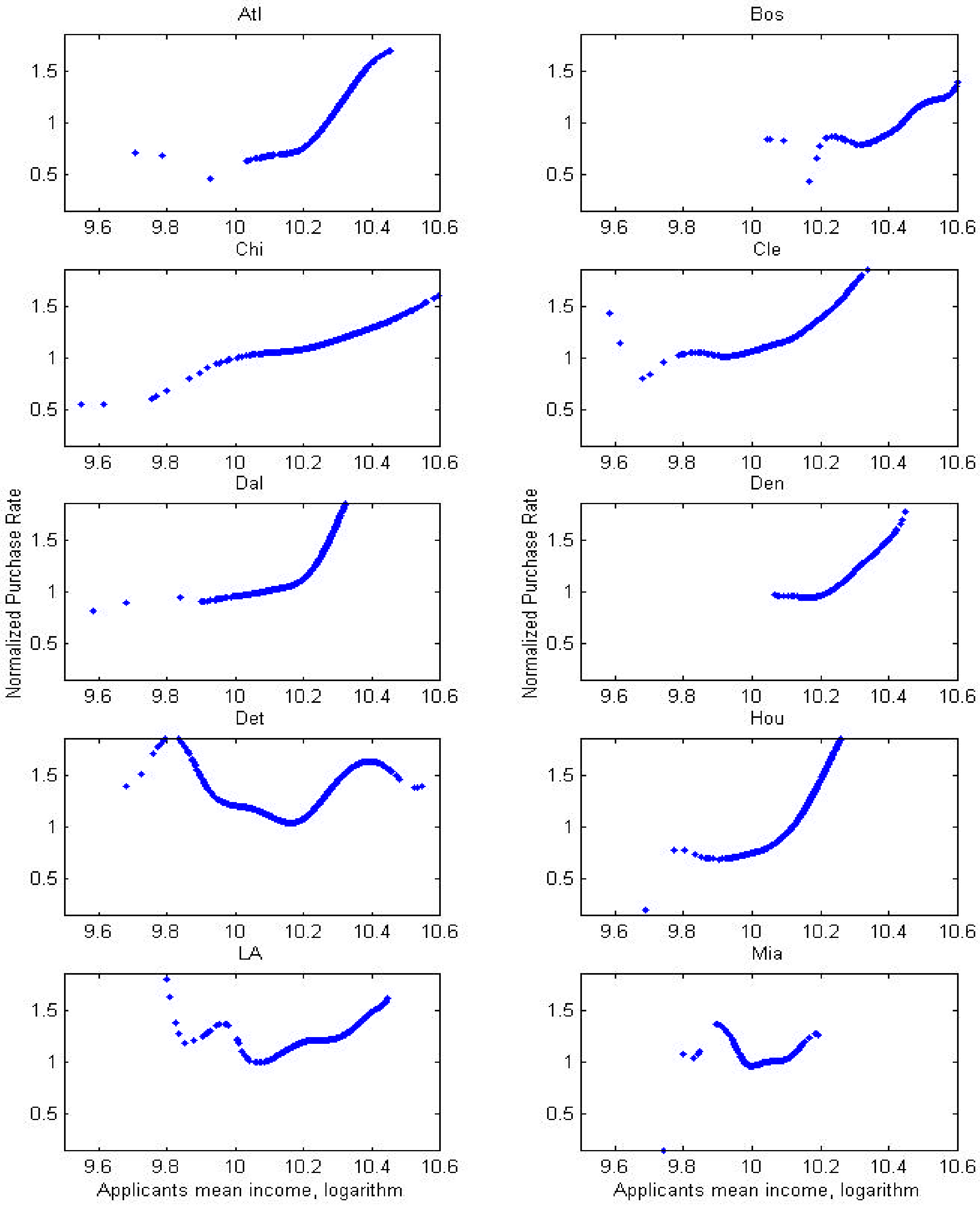


Figure A2. (continued) The non-linear relation of lower income applicants' tract mean LTV with estimated GSE purchase rate (NPR).


Atl
Bos
Chi
Cle
Dal
Den
Det
Hou
LA
Mia
Normalized Purchase Rate
Normalized Purchase Rate
0.5
1
1.5
9.6
9.8
10
10.2
10.4
10.6
Applicants mean income, logarithm
Applicants mean income, logarithm

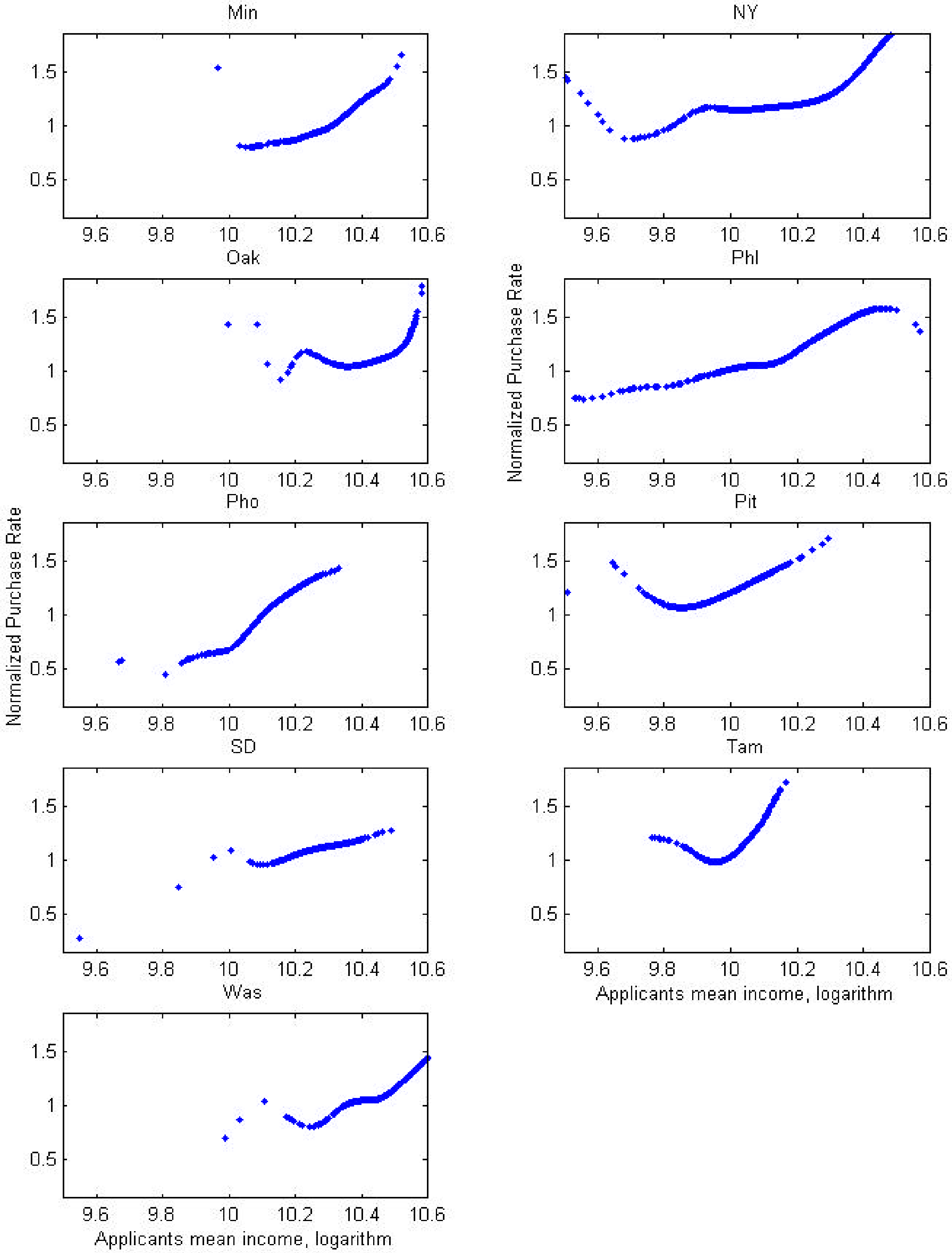

Min
NY
Oak
Phl
Pho
Pit
SD
Tam
Was
Normalized Purchase Rate
Normalized Purchase Rate
1.5
1
0.5
9.6
9.8
10
10.2
10.4
10.6
Applicants mean income, logarithm
Applicants mean income, logarithm

Figure A3. continued.