Spatial Variable Prediction of Housing Affordability in the Greater Toronto-Hamilton Area

by
William Russell

supervised by
Tarmo Remmel

A Major Paper
submitted to the Faculty of Environmental and Urban Change
in partial fulfilment of the requirements for the degree of Master in Environmental Studies
York University, Toronto, Ontario, Canada

August 21, 2026

Abstract

Housing affordability has become a notorious planning challenge in Canada, particularly within the Greater Toronto-Hamilton Area (GTHA), where rising housing costs have exacerbated inequalities. While existing research has extensively documented the political economy of housing and the consequences of neoliberal housing policy, fewer studies have applied machine learning techniques to identify the relative predictive power of socioeconomic, housing, accessibility, and land-use variables in explaining census tract-level affordability. This study addresses this gap through a spatial analysis of 2021 census tract data for the GTHA.

Using Geographic Information Systems (GIS), approximately 150 variables from Statistics Canada, the Canada Mortgage and Housing Corporation, municipal open data, and provincial datasets were compiled and processed into a unified spatial database. Classification and Regression Trees (CART) were used to explore the predictive relationships for housing affordability outcomes among variables before developing Random Forest regression models to predict the proportion of households living in affordable housing, defined using the Canada Mortgage and Housing Corporation's shelter-cost-to-income ratio (STIR) of below 30%. Getis-Ord G_i^* statistics were subsequently applied to the residuals of the final model to identify statistically significant spatial clusters where affordability outcomes remained unexplained.

The final Random Forest model explained approximately 79.7% of the variation in housing affordability across census tracts, demonstrating that housing affordability is strongly associated with interacting socioeconomic and housing-market characteristics rather than any single determinant. Variables relating to housing tenure, household composition, immigrant status, housing costs, subsidized housing, and building age emerged as the strongest predictors, whereas older housing builds and housing suitability categories contributed little to no predictive value. Residual hotspot analysis further revealed localized areas where unknown factors influence affordability beyond those captured by the model.

By integrating critical urban planning theory with explainable machine-learning methods, this research demonstrates how spatial data science can support planning practice while identifying structural patterns of housing inequality that warrant further policy attention.

Foreword

This Major Paper is the final component needed to satisfy the requirements of the Plan of Study for the Master in Environmental Studies (Planning) Program at York University. My Plan of Study's area of concentration is "Affordable Housing Planning and Policy", intersected by three components; affordable housing, equity planning, and public policy. My major paper is rooted in the first component of affordable housing, as an exploration of variable prediction for housing affordability in the Greater-Toronto Hamilton Area using spatial analysis. The other components, equity planning and public policy, are touched upon in the literature and reflected on during the discussion portion of the paper. This paper explores spatial disparities in housing affordability across the Greater Toronto-Hamilton Area using a plethora of variables and identifies pockets of data which goes unexplained by the underlying numbers. The combination of quantitative statistics and qualitative literature regarding housing affordability results in an interesting study that deepens the concepts explored in my Plan of Study.

Acknowledgements

I would like to begin by thanking my supervisor, Dr. Tarmo Remmel. His support, feedback, and encouragement during the process of writing this major paper has been unmatched. This has been an incredibly difficult year for Tarmo and his family, and I am eternally grateful for his continued support to make this paper become a reality despite the challenges he has faced.

I would also like to thank my advisor, Dr. Stefan Kipfer, for his guidance throughout the program and in shaping my Plan of Study.

Thank you to my family and friends for their love and support during my time in this program, I am extremely fortunate to have all of you in my life.

Table of Contents

Abstract ii
Foreword iii
Acknowledgements iv
Table of Contents v
List of Figures vi
List of Tables viii
1. Introduction 1
    1.1 The Three Cities Within Toronto 2
    1.2 Literature Review 5
    1.3 About Spatial Modelling 12
    1.4 Research Question and Objectives 18
2. Methods 19
    2.1 Data Preparation 21
    2.2 Classification & Regression Trees, Random Forest Modelling 26
    2.3 G_i* Statistics 31
    2.4 Data Gaps 31
3. Results 33
    3.1 Decision Trees & Random Forest Models 35
    3.2 Getis-Ord G_i* Statistics 63
4. Discussion 68
    4.1 Revisiting The Three Cities Within Toronto 68
    4.2 Inclusionary Zoning 70
    4.3 Revisiting the Literature 73
5. Conclusions 75
References 77

List of Figures

Figure 1. The Three Cities of Toronto (Hulchanski & University of Toronto, 2010, p. 2). 3
Figure 2. The Three Cities of Toronto, 2025 Projection (Hulchanski & University of Toronto, 2010, p. 27). 5
Figure 3. Example Classification Tree (Breiman et al., 1984, p. 21). 13
Figure 4. Splitting Classification Tree. 14
Figure 5. Example Regression Tree (Breiman et al., 1984, p. 219). 15
Figure 6. Percentage of Affordable Housing by Census Tract in the GTHA. 34
Figure 7. Decision Tree with all variables. 35
Figure 8. Decision Tree without affordability variables. 37
Figure 9. Decision Tree without core housing indicator variables. 39
Figure 10. Decision Tree without homeowners in core need. 41
Figure 11. Variable predictor importance ranking from the first random forest model. 43
Figure 12. Variable predictor importance ranking from the categorical-focused random forest model. 45
Figure 13. Variable predictor importance ranking from the random forest model with consolidation of vacancy rate. 47
Figure 14. Variable predictor importance ranking from the random forest model with trimmed housing completion variables. 49
Figure 15. Variable predictor importance ranking from the random forest model with road density instead of walkability. 51
Figure 16. Variable predictor importance ranking from the random forest model with population density as a continuous variable. 53
Figure 17. Variable predictor importance ranking from the random forest model with removing the remainder of core housing need variables. 55
Figure 18. Results of tuneRF Function to Demonstrate Optimal Value of mtry. 57
Figure 19. Variable predictor importance ranking from the random forest model using the optimal value of mtry. 58
Figure 20. Variable predictor importance ranking from the random forest model with trimmed vacancy and zoning variables. 60
Figure 21. Global Directional Effects Matrix of Variable Predictive Power on Housing Affordability. 62

Figure 22. Getis-Ord Gi* Hot Spot Analysis of Random Forest Residuals. 64
Figure 23. Distribution Profile of Random Forest Residuals. 65
Figure 24. Getis-Ord Gi* Threshold Distribution versus Actual Affordability Plot. 66
Figure 25. Median Income by Census Tract in Toronto Relative to Toronto CMA Average. 69
Figure 26. Major Transit Station Areas in Toronto CMA. 72

List of Tables

Table 1. Variables linked to theoretical concepts. 21
Table 2. Categorical representation of vacancy rate. 22
Table 3. Categorical representation of walkability. 24
Table 4. Categorical representation of population density. 25
Table 5. Comprehensive dataset of all continuous variables used in the analysis. 27
Table 6. Categorical variables converted as factor within R and appended to the primary dataset. 29
Table 7. Decision tree results using all variables. 36
Table 8. Decision Tree results without affordability variables. 38
Table 9. Decision Tree results without core housing indicator variables. 40
Table 10. Decision Tree results without homeowners in core need. 42

1. Introduction

Canada's housing affordability crisis is an exhaustive topic of debate. A long-term housing affordability study from the Wellesley Institute in the Toronto Census Metropolitan Area (CMA) indicates that affordability has broadly declined between the years of 1991 and 2016, with a distinct lack of progress being made towards rectifying preventable disparities in socio-economic demographics (Leon & Iveniuk, 2021). The definition of affordable housing can vary widely. The metric provided by the Canada Mortgage and Housing Corporation (CMHC), is the percentage of gross household income spent on shelter costs, including utilities (Leon & Iveniuk, 2021, p. 2). Based on this metric, the CMHC considers a household's shelter cost to be affordable if the shelter-to-income-ratio (STIR) is below 30% (Leon & Iveniuk, 2021, p. 2). There are criticisms associated with using STIR as a housing affordability measure, such as households that spend over 30% yet are not very financially stressed, which may be the case with high-income households, or those who spend more than 30% of their income on shelter by choice (Leon & Iveniuk, 2021, p. 2).

The findings from this study indicate that the number of households living in moderately unaffordable housing, defined as greater than 30% STIR but below 50%, rose from 203,640 in 1991 to 386,000 households in 2016, which is a 90% increase (Leon & Iveniuk, 2021, p. 4). In conjunction, the number of households living in severely unaffordable housing, defined as greater than 50% STIR, rose from 86,440 in 1991 to 205,940 households in 2016, which is a 138% increase (Leon & Iveniuk, 2021, p. 4). The authors identify 4 distinct policy recommendations to improve housing affordability equity, as follows (Leon & Iveniuk, 2021):

1. Introduce equity-based housing policy approaches, allowing for the elimination of disparities between populations;
2. Increase the supply of all housing types – social, affordable, and market housing – done through investment and reduced construction barriers to housing supply;
3. Tackle housing affordability through income support and rent benefits programs;
4. Reform tax subsidies to reduce housing affordability inequities and rebalance the support for renters and owners.

The aim of this study in spatial variable prediction of housing affordability in the Greater Toronto-Hamilton area (GTHA) is to investigate different socio-economic, housing, accessibility, and zoning characteristics to explain spatial disparities in housing affordability across the region.

As well, this study will identify which factors most strongly predict affordable housing outcomes as identified through Random Forest spatial modelling.

1.1 The Three Cities Within Toronto

In 2010, John David Hulchanski and the University of Toronto published *The Three Cities Within Toronto: Income Polarization Among Toronto's Neighbourhoods, 1970-2005*. This longitudinal study explored neighbourhood change in greater Toronto over a 35-year period (Hulchanski & University of Toronto, 2010). The study has many key findings with implications for housing policy. In the 35 years between 1970 and 2005, Toronto's neighbourhoods can be categorized by three distinct groups based on income change, coined the three cities (Hulchanski & University of Toronto, 2010, p. 1). City #1 is a high-income area where neighbourhood incomes rose greatly in comparison to the Toronto Census Metropolitan Area (CMA) average since 1970, and are concentrated in the central city and along subway lines. City #3 is a low-income area where neighbourhood incomes have fallen dramatically over the longitudinal study period relative to the CMA average, concentrated in the northeastern and northwestern parts of Toronto. City #2 is in between the two; which is a middle-income area where neighbourhood incomes have remained stable compared to the CMA average over 35 years. The most important finding of this report is a consistent, identifiable trend is observed by the researchers – the three neighbourhoods are changing at different rates and moving further apart (Hulchanski & University of Toronto, 2010, p. 1).

A significant observation is the dramatic proportionate shrinking of middle-income census tracts of the city, coinciding with a slight increase of high-income tracts, and a substantial increase in low-income tracts. The proportion of middle-income neighbourhoods, defined as incomes less than 20% above or below the CMA average, was 66% in 1970, but declined to 29% in 2005. Conversely, high-income neighbourhoods grew from 15% to 19%, while low-income neighbourhoods exploded from 19% to 53%. This trend can be observed across the rest of the Toronto CMA, which indicates that middle-income households are not simply moving to suburban municipalities outside of Toronto. The spatial distribution of income in Toronto has shifted over the study period as well. Most low-income neighbourhoods were found in the inner city in the 1970s, which allowed for superior transit and service access for these households. Due to gentrification, some of these neighbourhoods are now more affluent, which has caused the people forming low-income neighbourhoods to be redistributed out to the northeastern and northwestern parts of the

city, the inner suburbs, where transit and service access is poorer. Hulchanski's study identifies these patterns as persistent over the entire 35-year study period, where the city is polarizing into wealthy and disadvantaged neighbourhoods with a collapse of the middle class. Hulchanski emphasizes that the continued segregation of the city is not inevitable or irreversible, and calls for public policies to improve housing affordability, the expansion of access to transit and services in areas of greater need, and supporting the structural renewal of aging high-rise neighbourhoods found in City #3 (Hulchanski & University of Toronto, 2010, p. 1).

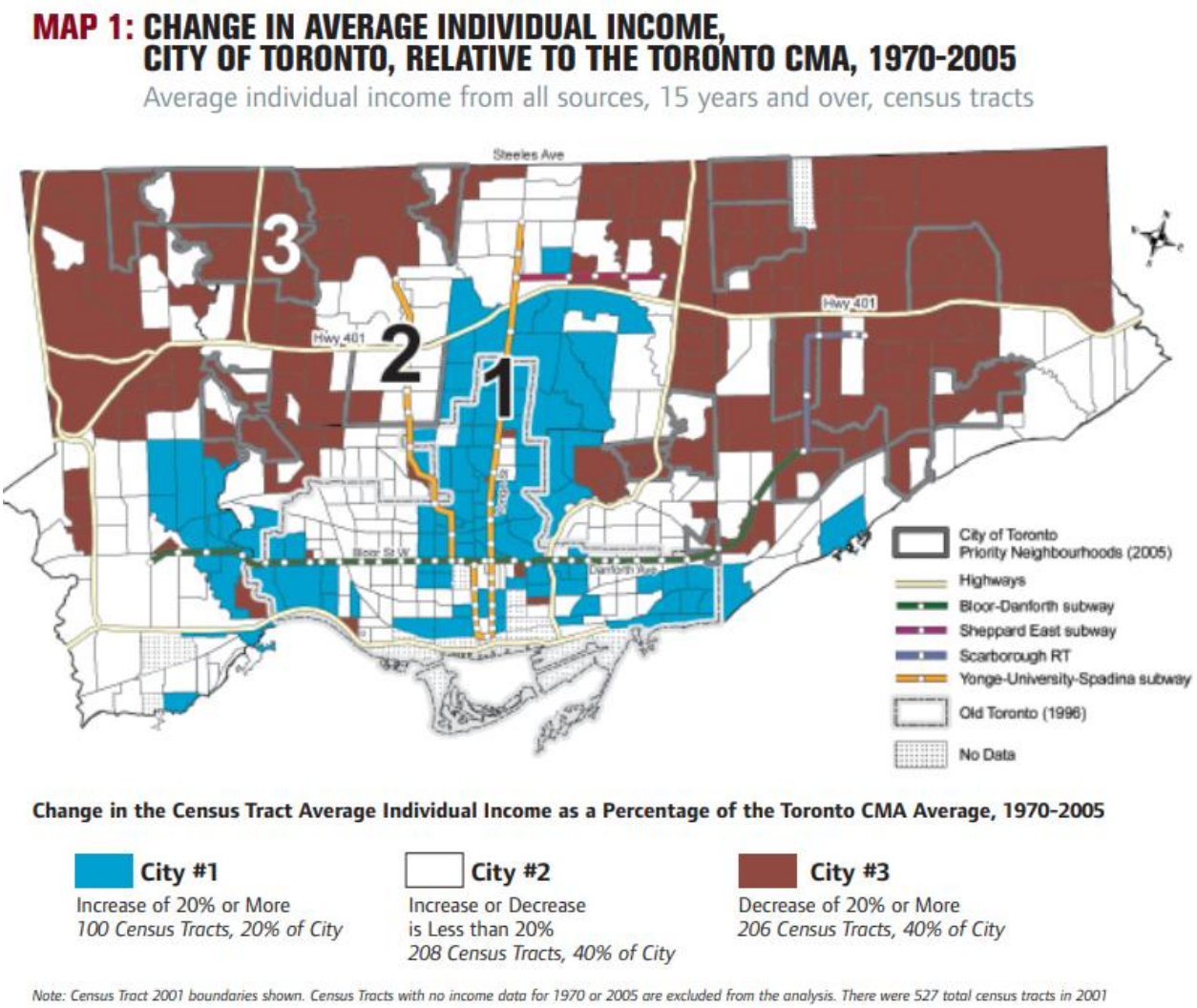


Figure 1. The Three Cities of Toronto (Hulchanski & University of Toronto, 2010, p. 2).

Hulchanski identifies the price of housing as a key determinant that leads to neighbourhood stability or change (Hulchanski & University of Toronto, 2010, p. 3). Since higher-income households can outbid lower-income households for preferred locations and housing quality, this can sometimes lead to shifting socioeconomic status at a neighbourhood scale. Sometimes, lower-income neighbourhoods may possess characteristics that are desirable to higher-income groups, which may cause gentrification and displace original residents. The opposite may also occur, if higher-income neighbourhoods fall out of favour and property values do not keep up with other neighbourhoods.

"Neighbourhood" is a loose definition, and ultimately, subjective. Neighbourhoods have tremendous impact on one's sense of community life, health, and educational outcomes in their overall well-being. Statistics Canada defines census tracts, which are considered "neighbourhood-like" local areas. Statistics Canada does this using recognizable physical boundaries, i.e., roads, railway lines, or rivers, defining compact shapes in which a more or less homogeneous population can be represented in terms of socio-economic characteristics (Hulchanski & University of Toronto, 2010, p. 4). In the study from Hulchanski and the University of Toronto, 531 census tracts are defined within the city of Toronto, with an average population of 4700 people, as of the 2006 census (Hulchanski & University of Toronto, 2010, p. 6). These numbers are drastically different from the study area of this paper, the Greater Toronto-Hamilton area, which consists of 1489 census tracts with an average population of 4867 people, as of the 2021 census.

Interestingly, Hulchanski's study concludes with a projection of the three cities in Toronto by 2025 if nothing changes. At its core, the prediction spells out a continuation of the trend analyzed over the 35-year study period, with a shrinking middle-class and expansion of both the higher- and lower-income cohorts. This projection is operating under the assumption of no major policy changes in the years since 2006, particularly policies that would contribute to a more equitable income distribution and an increase in the affordable housing supply (Hulchanski & University of Toronto, 2010, p. 27).

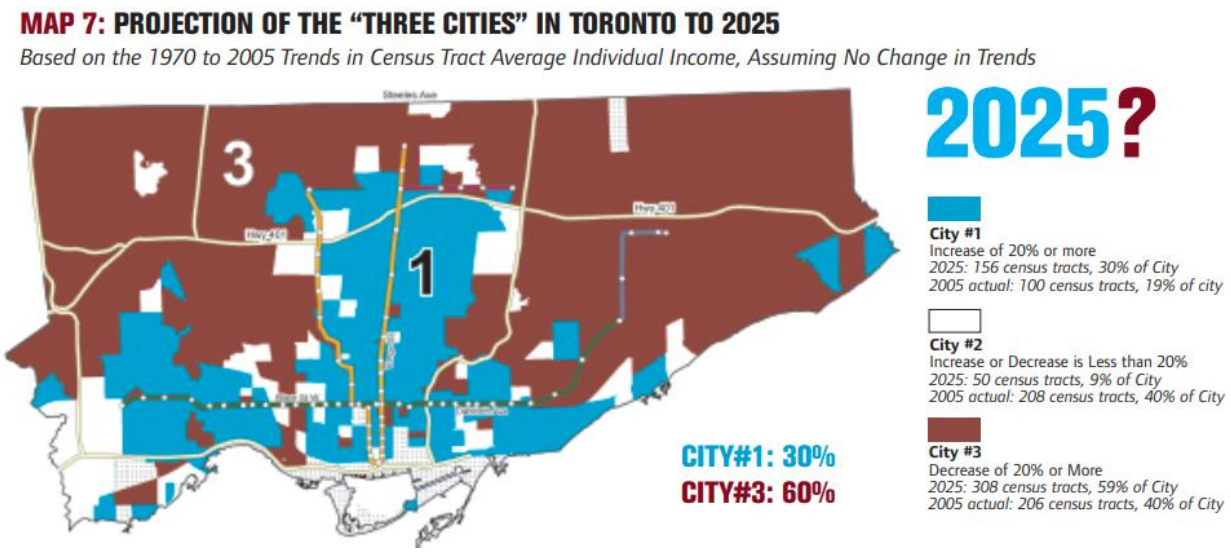


Figure 2. The Three Cities of Toronto, 2025 Projection (Hulchanski & University of Toronto, 2010, p. 27).

1.2 Literature Review

This study is grounded in a critical urban planning framework that views housing policy not only as a technical issue but also as a contested terrain shaped by political, economic, and social forces. It draws upon several interrelated bodies of literature, ranging from critical housing studies, spatial justice, and the political economy of urban governance.

In Defense of Housing provides a central foundation for the theoretical framework of this paper on spatial variable prediction of housing affordability (Madden & Marcuse, 2016). Madden & Marcuse (2016) seek not to defend the housing system, but the use of housing for inhabitants rather than speculative real estate (p. 11). Throughout much of human history, dwellings have been tied to broader social and economic relationships, with commodification of land being a very recent phenomenon (Madden & Marcuse, 2016, p. 18). There was a union between domesticity and labour described by Lewis Mumford, in the era of harsh feudal systems where peasants were tied to the land through their forced labour (Madden & Marcuse, 2016, p. 18). The precondition for such a transformation towards commodification is the privatization of the commons; thus, the systems surrounding communal regulation around housing and living must be destroyed (Madden & Marcuse, 2016, p. 18). A key theme regarding commodification is the alienation of residents, a term defined as not feeling at home, an experience defined by the current housing system by fear, stress, anxiety, and disempowerment (Madden & Marcuse, 2016, p. 54).

Hyper-commodified housing, housing used for perverse financial incentives, as investment vehicles, is alienating. It overwhelms the housing supply with individuals and capital investment firms that see dwellings as an opportunity for investment, rather than a social right (Madden & Marcuse, 2016, p. 58). Madden & Marcuse (2016) point out that luxury housing is antisocial, used by wealthy people to avoid encounters with poverty while contributing little to the community in which these buildings reside (pp. 37–38). Developers of such buildings pay little in tax and often receive massive public subsidies as an incentive for constructing less exclusive units, but poses inefficiencies when compared to the potential of what a nonprofit housing corporation could achieve given the same grants (Madden & Marcuse, 2016, p. 38).

The decline of alternatives to commodified housing is alarming; in New York, the regulated rental market share fell from 62% to 47% between 1981 and 2011, introducing precarity for tenants (Madden & Marcuse, 2016, p. 38). Many of these luxury buildings are not built to provide housing, but to make profits upon resale, propped up by exorbitant land values of areas considered to be prime real estate (Madden & Marcuse, 2016, p. 36). By engaging in the hyper commodification of housing, wealthy people have managed to monopolize more housing at the expense of the general population (Madden & Marcuse, 2016, p. 38). Gentrification has occurred at scale, where older buildings are targeted as "underperforming assets", renovated, and rent is demanded at a much higher rate than previously (Madden & Marcuse, 2016, p. 41).

David Harvey (2012) shares the critique of capitalist urbanization and market housing with Madden & Marcuse, although with a different focus (p. 6). As opposed to the idea of the alienation of housing causing social inequalities and oppression as purported by Madden & Marcuse, Harvey (2012) instead emphasizes the connection between housing and the circulation of surplus capital (p. 12). For Harvey (2012), the booms and busts of housing markets which occur as part of capitalism's cyclical dynamics are systemic outcomes driven by capital's perpetual need to find profitable investment in the search for wealth accumulation (p. 12).

Harvey (2012) unpacks this dynamic through the conceptualization of land not as an ordinary commodity, but as a form of "fictitious capital", a term used by Marx, which derives its value from an expected use of the land through future rent extraction (p. 10). This framework is enabled by neoliberal policies through the deregulation of property markets, relaxed zoning laws, and the desire for capital growth in urban spaces at the expense of human ecosystems (Harvey, 2012, p. 2). Commodification has driven low- and moderate- income households out of central

urban areas, observed in Manhattan and London, creating a lending cycle of fictitious capital that becomes “unreal estate”, with rampant speculative asset classes detached from the social utility housing should provide (Harvey, 2012, pp. 2, 11).

Furthermore, Harvey (2012) integrates the role played by financialization with reference to the credit system, and the part it plays regarding speculative capital of urban built environments (p. 10). Harvey (2012) notes the massive share of aggregate wealth held by residential mortgage markets, which exceed 40% of GDP in developed countries, this has a compounding macroeconomic effect on localized property disruptions (pp. 3, 6). This has immense consequences for ordinary people highly exposed to property value crashes, such as the housing crisis in 2007-09, where property owners and credit agencies engaged in a predatory practice, accumulation by dispossession (Harvey, 2012, pp. 2, 19). Disproportionately targeting African-American populations, some of these predatory practices include reverse redlining to persuade people into taking on predatory sub-prime loans as opposed to conventional, and fraudulent, illegal foreclosures on homes (Harvey, 2012, pp. 18–19). It is estimated that the low income African-American population of the United States lost between $71 and $93 billion in asset values due to predatory sub-prime loan practices during the housing crisis (Harvey, 2012, p. 18).

In *Social justice and the* city, Harvey (2009) emphasizes that questions of spatial justice link policy to urban form: redistributive housing policy must consider not only who benefits from housing, but also where these benefits are located and under what spatial conditions justice is materially realized (pp. 50–54). Strategies within urban systems contain and amend policies that are designed to change the city’s spatial form through objects such as housing, transport, employment zones, along with policies that affect social processes such as social structures, activities to connect people, and welfare systems (Harvey, 2009, p. 50). However, the harmonization of these policies with the overarching goal of a social objective poses an incredibly complex problem for planners, as there is a lack of a general consensus on what a coherent social objective entails (Harvey, 2009, p. 50). Harvey (2009) gives the example of prediction based on current knowledge and trends, future population distribution, consumption and travel patterns, to allocate investment accordingly, yet the very decision to invest based on current trends lends itself to the continuation of said trends (p. 51).

An example of investment feedback loops is the growth of automobile ownership and use in the United States, which can partly be attributed to investment policies that favoured roads and

highways compared to other transportation modes (Harvey, 2009, p. 51). Harvey (2009) points to the changing locations of economic activity in cities, thus relocation of job opportunities, along with changing residential activity leads to changing location of housing (p. 61). Both outcomes may be associated with changing expenditures on transportation, and can ultimately be tied to the occurrence of a wealth redistribution (Harvey, 2009, p. 61). Harvey (2009) identifies the pattern of low-income housing supply to be inelastic, a market failure of supply even when demand rises, and the location of stock is partially fixed by the existence of current housing stock available, and a strong social contiguity constraint (p. 61). It is for these reasons that the existence of low-income housing stock is primarily located in central city areas, and the slow reaction of demand for low-income housing supply in suburban areas (Harvey, 2009, p. 61). This issue is further complicated by attempts to expand the low-income housing supply in the inner city leading to undesirable outcomes such as higher prices for already poor quality housing, and incentivizes profit-seekers to commodify this housing supply further (Harvey, 2009, p. 61).

Kipfer and Sotomayor (2024) further argue that market-driven housing solutions often reinforce inequalities along racial, gendered, and class lines and stress the importance of anti-colonial analysis in understanding housing crises in settler-colonial contexts like Canada. Canada's settler-colonial context sheds light on the transformation of common land into privately owned real estate as it has been occurring since before Confederation (Kipfer & Sotomayor, 2024, p. 42). This land grab intentionally undermined the capacity of the population to sustain themselves independently of providing their labour for a wage, and gave substantial power to the Canadian ruling class surrounding finance, infrastructure, and extraction (Kipfer & Sotomayor, 2024, p. 42). This system fractured Indigenous connections to the land and interrupts land back initiatives as incompatible with capitalist property boundaries, blocking the possibility of "planning without property" models (Kipfer & Sotomayor, 2024, p. 42). The state-enforced landscape entrenched a residential hierarchy that consisted of male-controlled settler households and white-British supremacy (Kipfer & Sotomayor, 2024, p. 42).

Despite state investment in non-profit rental housing (social housing) during a short period beginning in the 1940s, followed by provincial-federal public housing policies in the 1960s and then cooperative housing programmes since the 1970s, the share of nonmarket housing in Canada has been shrinking in favour of a market housing approach (Kipfer & Sotomayor, 2024, pp. 42–43). This patriarchal and gendered system results in a disproportionate burden placed on women's access to

safe, affordable housing (e.g., domestic violence, gender pay gap, sexist urbanism) and the concealment of hidden homelessness of both women and members of the LGBTQ+ community (Kipfer & Sotomayor, 2024, p. 35). Working-class tenants, with a large proportion of non-white representation, are also disproportionately affected by the housing crisis (Kipfer & Sotomayor, 2024, p. 35).

It is these ways in which market housing solutionism reinforces inequalities, as the private market does not build based on societal need, but on profit maximization. The condo boom in Toronto, for example, has been gentrifying large public housing estates in the interest of individual owners and investor groups (Kipfer & Sotomayor, 2024, p. 43). It is flawed to assume the building of additional units proportional to projected population growth will solve the housing crisis, as what is currently being argued in Ontario (Kipfer & Sotomayor, 2024, p. 38). As such, the creation of new units tends to not benefit those who need it most, in contradiction to the trickle-down assumptions made in the housing supply argument (Kipfer & Sotomayor, 2024, p. 41). Housing affordability is a relational problem, an intersection of labour markets, state intervention, mobility, the environment, gender, and racialization (Kipfer & Sotomayor, 2024, p. 39).

Hackworth & Moriah (2006) demonstrate that since the 1990s, housing responsibility has devolved from the federal to municipal level without adequate fiscal resources (p. 515). This structural retreat began in 1993 with the federal government declaring that low-income housing is no longer a national responsibility (Hackworth & Moriah, 2006, p. 515). In Ontario, this restructuring accelerated under the provincial government led by Mike Harris, by cancelling all future housing commitments, eliminating support for existing units, and devolving all responsibility to municipalities (Hackworth & Moriah, 2006, p. 515).

The restructuring came with a multitude of consequences. The first and most obvious outcome was an intensified affordable housing crisis, particularly in urbanized provinces like Ontario, where rates of homelessness for Canadians exceeded the United States on a per capita basis in 2000 (Hackworth & Moriah, 2006, p. 516). The second outcome was an unnecessarily difficult-to-navigate landscape of social housing provision alongside a decrease of expertise in the area due to the dramatic restructuring, with many complaints from both private and municipal non-profit providers (Hackworth & Moriah, 2006, p. 516). The third outcome, largely unexplored at the time of this article's publication, is the opportunity for experimentation from primarily non-profit

housing providers and municipalities to create more affordable housing, but also holds potential dangers for further privatization of housing stock (Hackworth & Moriah, 2006, p. 516).

More recent work by August and Walks (2018) illustrates how neoliberal policy shifts to withdraw from social housing provisions have contributed to gentrification and displacement, in inner urban districts (p. 134). There was a shift towards reinvestment in Toronto's multi-family housing stock after decades of decline, enabled by neoliberalism to deregulate rental protections and decontrol rent, which came with its own set of problems with financialization (August & Walks, 2018, p. 134). The financial penetration into Toronto's multi-family housing stock came about from real estate investment trusts (REITs), which began as a response to favourable market conditions facilitated by the aftermath of the global financial crisis (August & Walks, 2018, p. 134). The resulting financial boon for investors came at the expense of renters in the form of a rental affordability crisis (August & Walks, 2018, p. 127).

Equity planning, as developed by Krumholz (1986) and extended by other scholars, described by Krumholz and Hexter (2019), offers a normative planning framework aligned with the goals of this study. Krumholz (1986) notes that equity planners should aim to distribute resources, opportunities, and services more equitably across urban populations to the benefit of low-income, disadvantaged groups (p. 331). Krumholz (1986) draws on his experience as the city planning director for Cleveland in the 1970s to achieve equitable planning solutions for housing, transportation, and service improvements for disadvantaged groups (p. 331). Krumholz (1986) identifies the importance of citizens benefitting from granting subsidies to private developers, and the importance of performing cost/benefit analysis on projects where the intention was to benefit a targeted group, if this was not possible, the project was scrapped (p. 330). Krumholz (1986) is also highly supportive of investment into community organizations, believing that helping local neighbourhood-based organizations execute their own agendas is an important part of equity planning (1986, p. 334).

According to Krumholz & Hexter (2019), one of the most substantial contributors to equity planning is Paul Davidoff. Often described as the key exponent of advocacy planning, Davidoff stressed the merit of developing alternative plans to advocate for social groups typically excluded from institutional planning processes, and, ultimately, facilitate genuine deliberation in these processes. While these foundational contributions define the moral core of equity planning, scholars have questioned their limitations. For instance, Zapata & Bates (2015) argue that much of

the literature dominated by white male planners have neglected how racialized and gendered inequalities shape access to urban resources. They call for a renewed, intersectional approach that centers communities of colour and re-grounds planning practice in racial and social justice (Zapata & Bates, 2015, pp. 245–248). A few of their lessons for a renewed equity planning approach include the idea that topics that are important to communities of colour should matter to planners, to be an effective equity planner requires a shared goal and vision, and the necessity for successful equity planning to be rooted in successful political and governance strategies (Zapata & Bates, 2015, p. 247).

This framework remains highly relevant in Toronto, where planners operate within institutional constraints shaped by neoliberal policies. Lehrer et al. (2010) demonstrate that urban redevelopment projects such as Regent Park's "revitalization" were justified through equity rhetoric but in practice reproduced gentrification and displacement. The learnings from the regent park model inspire hope in a call for investment into real affordable housing projects which do not result in a privatized "cleansing operation" which renders racialized populations invisible (Lehrer et al., 2010, p. 88). Similarly, Whitzman & Slater (2006) show how local activism in Parkdale was undermined by the same market forces that equity planning sought to resist. The terms "ghetto" and "village" are labels which legitimize gentrification, with private units wrought with neglect, and current tenants ignored by policymakers (Whitzman & Slater, 2006, p. 693). These studies highlight the enduring tension between planning's redistributive ambitions and neoliberal governance, suggesting the need for new methods to empirically trace how inequality is embedded in urban form.

From a public policy standpoint, Marcuse (1978) challenges the assumption that the state is acts to promote the welfare of its citizens as its primary goal, linking this idea with housing policy (p. 36). He argues that housing policy is often designed to serve the wealthy elite, citing that state support for housing provision for poorer families was not widespread until it was in the interest of private profit to do so (Marcuse, 1978, p. 41). Such policies include housing regulation and the public provision of housing in New York City (Marcuse, 1978, pp. 38–39). He begins with the notion that housing provisions were not done out of concern for citizen welfare, but to serve existing capital interests of labour, and reduction in class and ethnic conflict (Marcuse, 1978, p. 38). Part of the motivation for public provision of housing in the United States was to support the war effort in the First World War, rather than concern over discontent from returning war veterans (Marcuse,

1978, p. 39). This is furthered by the motivation for the 1937 public housing program being led by the provision of jobs over the provision of housing, to serve capital interests (Marcuse, 1978, p. 40). The largest opposers of public housing provision in the United States at the time were the National Association of Real Estate Boards, the United States Savings and Loan League, and Mortgage Bankers of America (Marcuse, 1978, p. 40). This was due to the view that these groups favoured the principle of urban redevelopment as a way to enable private capital to do the job instead of public housing (Marcuse, 1978, p. 40). This understanding aligns with Kipfer & Sotomayor's argument that contemporary housing policy in Canada reflects "market housing solutionism", wherein state actors reproduce the very inequalities they claim to address by prioritizing supply-driven market interventions (Kipfer & Sotomayor, 2024, p. 37).

It is important to note that census tracts, the statistical data boundary used for most of the data collection in this study, are inherently arbitrary. Census tracts are statistical boundaries defined by Statistics Canada, divided with boundaries to resemble pockets of the population as it is digestible at that moment in time. As people move and the urban built form evolves, so do census tracts, which may result in shifting boundaries from census to census. It is also important to remember the colonial history of Canada, leading to severed ties between Indigenous people and their ancestral lands, and its historically patriarchal housing system which preceded the country's Confederation. These are important historical contexts to keep in mind when using arbitrarily defined statistical boundary data provided by the Canadian government.

1.3 About Spatial Modelling

This research undertakes this question using a quantitative geospatial data-science approach. Geographic Information Systems (GIS) are integrated with machine-learning classification and regression, applying a Random Forest (RF) model to identify nonlinear relationships among predictors of affordability. Random Forest is a decision-tree based method which reduces user bias, manages complex interactions between variables, and yields interpretable variable-importance rankings (Friedl & Brodley, 1997).

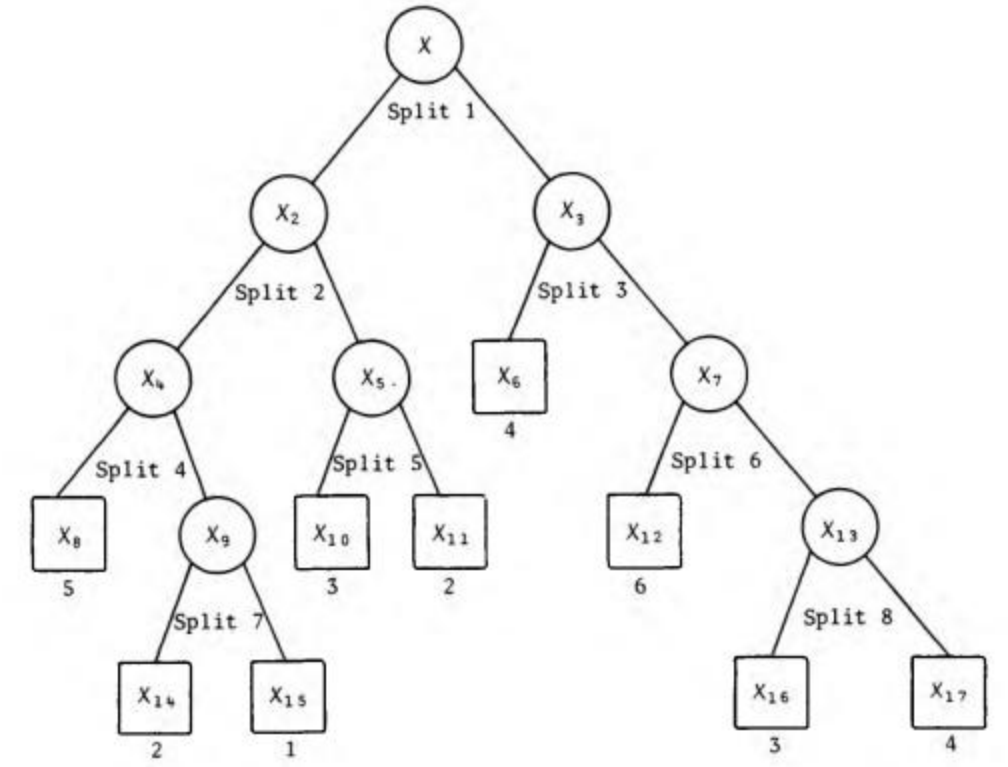


Figure 3. Example Classification Tree (Breiman et al., 1984, p. 21).

Figure 3 is an example of a classification tree from chapter 2 of the book *Classification and Regression Trees* (Breiman et al., 1984, p. 21). The model begins with the complete dataset at a single root node and repeatedly divides the observations into two groups based on the predictor that best separates the data. Each split creates two child nodes, which may be split again if doing so improves the classification. This recursive process continues until no further meaningful splits can be made or another stopping criterion is reached. The nodes at the end of each branch, known as terminal (or leaf) nodes, contain the final classifications or predictions produced by the tree, while the intermediate nodes represent decision points where the data are partitioned according to the selected predictor variable. Figure 4 visualizes this concept, where the split is made such that the child subsets are divided equally. This pattern is continued in further splits for nodes t_2 and t_3, each descendant node ideally more uniform than the last, until information from the data set is exhausted.

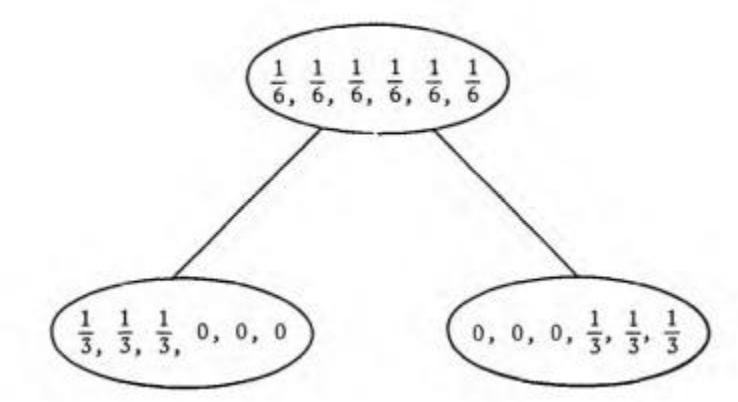


Figure 4. Splitting Classification Tree.

Figures 3 and 4 illustrate the general structure of a classification tree. Regression trees follow the same recursive splitting process but differ in the type of outcome they predict. Rather than assigning observations to discrete categories, a regression tree predicts a continuous numerical value. Each internal node contains a decision rule based on a predictor variable, while each terminal node contains the predicted value for all observations that reach that point in the tree. The prediction at a terminal node therefore reflects the combined effect of every splitting decision made along the path from the root node.

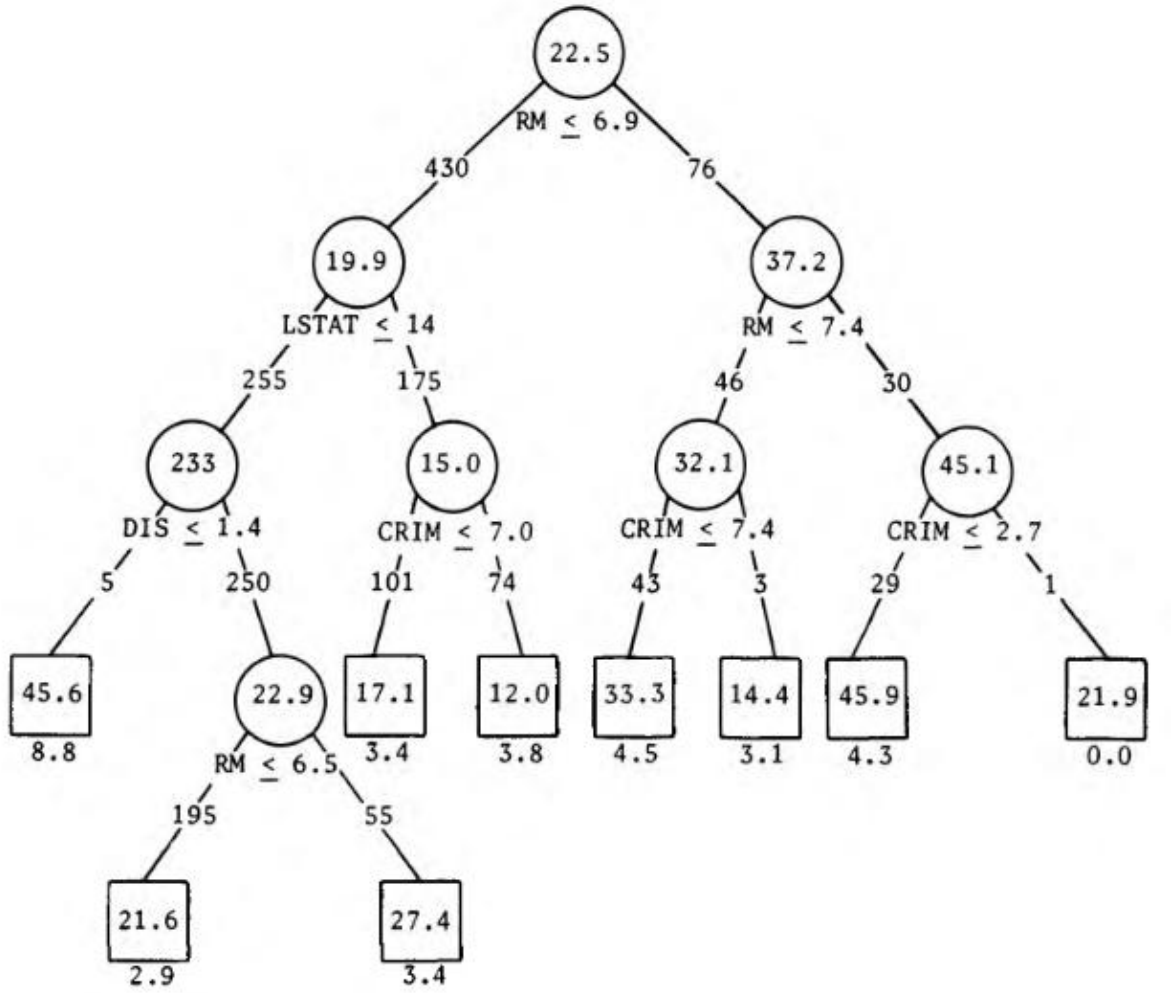


Figure 5. Example Regression Tree (Breiman et al., 1984, p. 219).

Figure 5 demonstrates an example regression tree, used in a 1978 paper, *Hedonic prices and the demand for clean air* by Harrison and Rubinfeld, to analyse data about Boston housing values by census tract with the following variables:
(Breiman et al., 1984, p. 217)

- y: median value of homes in thousands of dollars (MV)
- x1: crime rate (CRIM)
- x2: percent land zoned for lots (ZN)
- x3: percent nonretail business (INDUS)
- x4: 1 if on Charles River, 0 otherwise (CHAS)
- x5: nitrogen oxide concentration, pphm (NOX)

- x6: average number of rooms (RM)
- x7: percent built before 1940 (AGE)
- x8: weighted distance to employment centers (DIS)
- x9: accessibility to radial highways (RAD)
- x10: tax rate (TAX)
- x11: pupil/teacher ratio (P/T)
- x12: percent black (B)
- x13: percent lower-status population (LSTAT)

Each node contains the average of the dependent variable (MV) over all census tracts contained within the node. For example, the parent node shows 22.5, as the average value for MV across all census tracts. The tree is then split based on a decision made by the regression model, such as RM ≤ 6.9, where all values that satisfy the condition are sent to the left, while all values that do not satisfy the condition are sent to the right. In this case, 76 census tracts do not have an average number of rooms (RV) less than or equal to 6.9, while 430 census tracts do. As the tree reaches the terminal nodes, outlined by squares, they are denoted by the predicted value of MV for all census tracts that followed each respective branch. The number below each terminal node represents the estimated standard deviation of MV within the given terminal node, where lower numbers mean a tighter prediction (Breiman et al., 1984, pp. 217–220). Not every variable shows up as a decision point on a regression tree. In this case, only RM, LSTAT, DIS, and CRIM showed enough predictive power to warrant a decision being made to split the nodes.

Random Forest models take classification and regression trees to an extreme, allowing for accurate predictions less prone to data overfitting (Prasad et al., 2006, p. 184). Random forest models allow for a specified number of trees to be generated and creates splits at random intervals rather than trying to select an optimum split at each node. The reason for this method is that classification and regression trees (CART) cannot see up/down the tree beyond the current decision node. As a result, there is no way of knowing whether a different, less optimal decision now will produce a better split later. Random Forest navigates this problem by randomization and replication rather than the instantaneous optimization seen in CART. The trees within a random forest model are grown to maximum size using a bootstrap sample drawn from the original dataset (Prasad et al., 2006, p. 184). Each bootstrap sample leaves out approximately one-third of the observations, referred to as out-of-bag (OOB) samples. After a tree is constructed, its OOB samples

are used as an independent validation dataset to estimate prediction error and calculate variable importance without requiring a separate testing dataset (Prasad et al., 2006, p. 184). A caveat to random forest is the inability to analyse individual trees like with one classification or regression tree, however, this can be overcome through variable importance calculation visualizations along with the other benefits it provides through pure data aggregation.

This study uses the Getis-Ord G_i^* statistic (Gi-star), which measures the degree of association from the area represented by a weighted point, and all other weighted points within a distance (Getis & Ord, 1992, p. 190).

$$G_i(d) = \frac{\sum_{j=1}^{n} w_{ij}\,(d)x_j}{\sum_{j=1}^{n} x_j}, \quad j \text{ not equal to } i,$$

- $G_i(d)$ (The statistic)
    - A local distance statistic measuring the spatial concentration of a positive variable *X* within a given radius *d* around a specific point *i*, explicitly incorporating the value of point *i* itself (Getis & Ord, 1992, p. 192).
- *i* (Focal point)
    - The target region or point of interest being evaluated for spatial association (Getis & Ord, 1992, p. 190).
- *j* (Neighbouring points)
    - The points or regions distributed across the study area being checked against the focal point *i* (Getis & Ord, 1992, p. 190).
- *n* (Total number of regions)
    - The total count of mapped points comprising the full geographic scope of the study (Getis & Ord, 1992, p. 190).
- *d* (The radius/distance)
    - The specified threshold distance defining the physical spatial boundary around point *i* (Getis & Ord, 1992, p. 190).

The G_i^* Statistic is a powerful tool for identifying broader tendencies of positive spatial clustering and locating specific local "pockets" of high and low spatial association (Getis & Ord, 1992, p. 201). To compute this, Getis and Ord (1992) establish an area subdivided into *n* distinct regions, where *i* = 1, 2, ..., *n*, and each region is identified by known Cartesian coordinates (p. 190).

Every location *i* is associated with a positive attribute value *x* taken from a variable *X* that possesses a natural origin (Getis & Ord, 1992, p. 190). The derived formula tests statistical hypotheses regarding the spatial concentration of the sum of these *x* values across all *j* points located within a specific distance threshold *d* of the target point *i* (Getis & Ord, 1992, p. 190).

Unlike the standard G_i formula, the spatial weight matrix $w_{ij}(d)$ for the G_i^* variant allows the target point to be included in its own neighbourhood, meaning *j* may equal *i* (Getis & Ord, 1992, p. 192).

Quantitative modelling is unable to capture the full complexity of housing as a lived experience. However, it can expose the systemic patterns through which inequality is produced and sustained. By deploying the methodology used under a critical framework, the study contributes to an emergent field of socially conscious spatial data science, with commitment to housing justice.

1.4 Research Question and Objectives

This study poses the following research question:

How do socioeconomic, housing, accessibility, and land-use characteristics interact to explain spatial disparities in housing affordability across the Greater Toronto-Hamilton Area, and which factors most strongly predict census tract-level affordability using Random Forest spatial modelling?

This study has four primary objectives:

1. To identify the socioeconomic, housing, accessibility, and land-use characteristics associated with spatial variation in housing affordability across the GTHA.
2. To develop and evaluate Random Forest regression models capable of predicting census tract-level housing affordability.
3. To assess the relative importance and direction of influence of predictor variables in explaining housing affordability outcomes.
4. To identify statistically significant spatial clusters of model residuals using Getis-Ord G_i^* statistics to highlight areas where affordability is influenced by factors beyond those represented in the model and discuss their implications for housing policy and planning.

2. Methods

2021 Census data from Statistics Canada, obtained using the Canadian Census Analyser from Information & Institutional Technology from the University of Toronto, is the backbone of this study (Statistics Canada, 2023a; University of Toronto, 2022). The following lines of data from the 2021 census for each of the Toronto, Oshawa, and Hamilton census metropolitan areas were used in this analysis:

- Median total income in 2020 among recipients ($); Both sexes
- Average household size
- Married; Both sexes
- Living common-law; Both sexes
- Not married and not living common-law; Both sexes
- Non-immigrants; Both sexes
- Immigrants; Both sexes
- Immigrants; Both sexes / 2011 to 2021
- Private households by tenure — 25% sample data / Owner
- Private households by tenure — 25% sample data / Renter
- Dwelling provided by the local government, First Nation or Indian band
- Occupied private dwellings by period of construction — 25% sample data
    - 1960 or before
    - 1961 to 1980
    - 1981 to 1990
    - 1991 to 2000
    - 2001 to 2005
    - 2006 to 2010
    - 2011 to 2015
    - 2016 to 2021
- Occupied private dwellings by dwelling condition — 25% sample data
    - Major repairs needed
- Private households by housing suitability — 25% sample data
    - Suitable
    - Not suitable

- Owner and tenant households with household total income greater than zero, in non-farm, non-reserve private dwellings by shelter-cost-to-income ratio — 25% sample data
    - Spending less than 30% of income on shelter costs
    - Spending 30% or more of income on shelter costs / 30% to less than 100%
- Total — Occupied private dwellings by housing indicators — 25% sample data
    - Households 'spending 30% or more of income on shelter costs' or 'not suitable' or 'major repairs needed'
        - Spending 30% or more of income on shelter costs only
        - Not suitable only
        - Major repairs needed only
    - 'Spending 30% or more of income on shelter costs' and 'not suitable' and 'major repairs needed'
    - Acceptable housing
- Total — Owner and tenant households with household total income greater than zero and shelter-cost-to-income ratio less than 100%, in non-farm, non-reserve private dwellings — 25% sample data
    - In core need
    - Not in core need
- Total — Owner households in non-farm, non-reserve private dwellings — 25% sample data
    - % of owner households with a mortgage
    - % of owner households spending 30% or more of its income on shelter costs
    - % in core housing need
    - Median monthly shelter costs for owned dwellings ($)
    - Median value of dwellings ($)
- Total — Tenant households in non-farm, non-reserve private dwellings — 25% sample data
    - % of tenant households in subsidized housing
    - % of tenant households spending 30% or more of its income on shelter costs
    - % in core housing need
    - Median monthly shelter costs for rented dwellings ($)
- Households living in a dwelling provided by the local government, First Nation or Indian band in non-farm private dwellings — 25% sample data

    - % of households living in a dwelling provided by the local government, First Nation or Indian band spending more than 30% on shelter costs
    - Median monthly shelter costs for dwellings provided by local government, First Nation or Indian band ($)
- Unemployment rate; Both sexes

In the planning phase of this study, numerous potential variables were identified with respect to theoretical concepts discussed in the literature review:

Table 1. Variables linked to theoretical concepts.

Variable	Theoretical Representation
Income	Class struggle, purchasing power
Household composition	Social reproduction, care burden
Newcomer share	Racialized & migration-linked vulnerability
Tenure	Decommodification, market delivery
Building age	Filtering vs reinvestment, gentrification
Vacancy rate	Market dynamics, commodification
Rent level	Direct market outcome
Market sector status	(De)commodification, policy outcomes
Transit access	State-led spatial capital investment
Job access	Spatial mismatch, economic exclusion
Walkability	Amenity-led rent escalation
Density permissions	Zoning as supply regulation
Inclusionary zoning	Redistributive planning intervention
Building permits	Market responsiveness, speculative cycles

2.1 Data Preparation

The data preparation for this study involved a few moving parts to patch it all together. The first thing to settle was administrative boundary selection. The 2021 Census from Statistics Canada provides the framework upon which other data parts were stitched together. The raw numbers for approximately ~80 different variables from the 2021 Census were compiled using the Computing in the Humanities and Social Sciences (CHASS) Canadian Census Analyser from the University of Toronto, set for the Census Metropolitan Areas of Toronto, Hamilton, and Oshawa, which covers the entirety of the Greater Toronto-Hamilton Area. The granularity of data chosen were census tracts, due to the uniformity of boundaries and with the idea of a finer scope providing more potential for data variability, thus resulting in a more interesting study. Statistics Canada has digital boundary

files for all administrative and statistical boundaries, which can be used to join the raw numbers for these census tracts to its corresponding spatial boundary in a Geographic Information System. In this case, ArcGIS Pro was used for all spatial data processing.

There are a few limitations of the census data amongst the study variables. This includes Vacancy Rate, Transit Access, Walkability, Density Permissions, Inclusionary Zoning, and Building Permits. Each of these variables had varying degrees of data availability, ranging from some being limited through data not being fine enough, to some requiring an estimation via a proxy variable.

Vacancy Rate data was obtained through Canada Mortgage and Housing Corporation (CMHC), a crown corporation which partners with Statistics Canada to deliver accurate housing data. This variable was obtained at the census tract level for the annualized year 2021, to match as closely as possible to the census year 2021. Vacancy Rate lacked some data at the census tract level, with many tracts either not having data at all, or some data being made unavailable for the public. This is due to the variable's focus on a broader geographic scope, and in some cases, identity protection can be a factor. Vacancy rate was later replicated as a categorical variable due to a lack of data being represented by "**" in many of the census tracts. The following table demonstrates how vacancy rate was categorized:

Table 2. Categorical representation of vacancy rate.

Category	**Threshold**
N/A	** (No data)
Vacancy Rate Studio - Zero	0
Vacancy Rate Studio - Low	≥0%, <3.0%
Vacancy Rate Studio - Medium	≥3.0%, <7.0%
Vacancy Rate Studio - High	≥7.0%
Vacancy Rate 1 Bedroom - Low	≥0%, <2.5%
Vacancy Rate 1 Bedroom - Medium	≥2.5%, <5.0%
Vacancy Rate 1 Bedroom - High	≥5.0%
Vacancy Rate 2 Bedroom - Low	≥0%, <2.0%
Vacancy Rate 2 Bedroom - Medium	≥2.0%, <4.0%
Vacancy Rate 2 Bedroom - High	≥4.0%
Vacancy Rate 3+ Bedroom - Low	≥0%, <1.5%
Vacancy Rate 3+ Bedroom - Medium	≥1.5%, <4.0%
Vacancy Rate 3+ Bedroom - High	≥4.0%
Vacancy Rate Total - Low	≥0%, <2.0%
Vacancy Rate Total - Medium	≥2.0%, <4.0%
Vacancy Rate Total - High	≥4.0%

It was important to categorize vacancy rate by apartment size individually based on the extreme variance in data, to normalize the distribution between census tracts receiving classifications of low, medium, and high vacancy rates. Among the 1489 census tracts analyzed in this study, only 227 had vacancy rate data at all. As such, Null was a large category. When dealing with Null values in a decision tree in R, another possible solution is to populate missing values with the median of each column (R Core Team, 2026). Creating an N/A category was preferable to this option, due to the large number of census tracts without reliable data, having most of the data be extrapolated from a small dataset would introduce too much noise into a random forest model without a way to reliably tell if vacancy rate was influencing the independent variable legitimately or through the extrapolated median.

Transit Access is an example of a proxy variable being used to identify the access to transit by census tract. The census questionnaire does not go into detail regarding transit access beyond the Journey to Work category, which can be related but is not necessarily indicative of the transit accessibility in each census tract. To combat this, a proxy variable for proximity to transit stations was used. Municipalities, and in some cases, regions, in the GTHA, as well as the provincial government, have transit data sets available for public use. This can be in the form of geographic feature layers such as shapefiles or geoJSONs, or as real time General Transit Feed Specification (GTFS) services. All transit stops for all forms of transportation (i.e., bus, train, subway), were overlaid on the census tracts in ArcGIS Pro, obtained from the following sources: (City of Brampton, 2025; City of Burlington, 2026; City of Hamilton, 2026b; City of Mississauga, 2026; Government of Ontario, 2026a; Metrolinx, 2026; Regional Municipality of Durham, 2020; The Regional Municipality of York, 2024; Toronto Transit Commission, 2026; Town of Milton, 2024).

The goal was to create a proxy that measures the general access to transit by counting the number of transit stops likely accessible to a given household present in each census tract. ArcGIS Pro geoprocessing tools were used to achieve this, where the centroid point of each census tract polygon was identified and a 1 km circular buffer was calculated for each of these centroid points. A new field “Transit Score” was calculated by summarizing the absolute count of transit stops that were physically located within each 1 km buffer, to give a proxy score of transit accessibility for each census tract. This poses a limitation in the study, as the proxy for transit accessibility assumes equivalency in transit time/cost/effort in all directions, due to the transit score calculation being the sum of all transit stops within a circular buffer.

Walkability is more difficult to measure. Like Transit Access, the census gives information about the number of people that walk to work, but this is not a reliable indicator of general walkability in a neighbourhood. To combat this problem, another proxy variable was used in the form of road density for Local/Street level road segments as defined by the Ontario Road Network (Government of Ontario, 2026b). The Ontario Government maintains an authoritative spatial data set detailing information about the road network in Ontario, of which only local/street level road segments were extracted as a proxy for local walkability in each census tract. Road density was measured in kilometre per square kilometre (km/km^2), taking the total street level road length per square kilometre of the census tract. Road density was then categorised into a variable labelled walkability by the following thresholds:

Table 3. Categorical representation of walkability.

Category	**Threshold**
Low Walkability	<4 Road density (km/km^2)
Medium Walkability	≥4, <10 Road density (km/km^2)
High Walkability	≥10 Road density (km/km^2)

Density permissions were difficult to quantify. Upon inspection of the official plans and zoning by-laws for every municipality in the GTHA, it was determined that there was a lack of consistency present for defined density permissions that would allow for meaningful data analysis. Most municipalities maintain publicly usable spatial data sets for their zoning by-laws with varying degrees of recency, although a few smaller ones (Halton Hills, Caledon, Uxbridge, Pickering) do not. In these cases, their zoning by-law maps are either only viewable in pdf format divided into sections, or any interactive online mapping tool available has its data barred from public use. In these cases, it can be extremely difficult to decipher the zoning codes applied on a per-property basis. Due to lack of data cohesiveness, population density from the 2021 census was used in place of analyzing density permissions from zoning by-laws. Population density was represented in two ways; a continuous variable using the value provided by Statistics Canada for each census tract, and; a categorical variable grouped by the following thresholds:

Table 4. Categorical representation of population density.

Category	Threshold
Low Density	<2500 people/km^2
Medium Density	≥2500, <5500 people/km^2
High Density	≥5500, <15500 people/km^2
Ultra-High Density	≥15500 people/km^2
Rural	Majority of land use classified as Rural

Classifying rural areas for population density purposes is what the spatial data obtained from zoning by-laws was ultimately used for. Every municipality classifies land use differently, but an effort was made to consolidate land use into eight distinct categories:

1. Commercial
2. Employment/Industrial
3. Infrastructure/Utilities
4. Institutional
5. Mixed Use
6. Open Space/Conservation
7. Residential
8. Rural/Agricultural

Doing so required some judgement calls on specific municipalities' definition of land use, and required manual classification by looking at land use schedule maps in cases where spatial data was not available for public use or available at all. Each census tract was then assigned one zoning category, based on which one occupied the highest percentage of land area in each census tract. This allowed for a new variable, Zoning, to be used in the proceeding analysis, alongside population density as a continuous or categorical variable. The following sources were used, which include spatial datasets and zoning by-law documents (*Caledon Zoning By-Law*, 2025; *Oakville Zoning By Law 2009-189*, 2024; *Oakville Zoning By Law 2014-014*, 2024; City of Brampton, 2026; City of Burlington, 2023; City of Hamilton, 2026a; City of Markham, 2024; City of Mississauga, 2021; City of Oshawa, 2017; City of Pickering, 2024; City of Richmond Hill, 2026; City of Toronto, 2024; City of Vaughan, 2026; Municipality of Clarington, 2026a, 2026b; Town of Ajax, 2021; Town of Aurora, 2025; Town of East Gwillimbury, 2025; Town of Georgina, 2021; Town of Halton Hills, 2025;

Town of Milton, 2022, 2026; Town of Uxbridge, 2023; Town of Whitby, 2024; Town of Whitchurch-Stouffville, 2026; Township of King, 2025).

Building permits did not materialize in the way it was initially envisioned, but it was ultimately replaced by Housing Completions data from CMHC. CMHC publishes monthly and annualized housing data in a variety of statistical geographies, including census tracts. Housing completion data was obtained on a census tract scale for the GTHA, annualized to the year 2021 to more closely align with the 2021 census data (CMHC, 2026). CMHC also splits this data in two ways; first by dwelling types (single, semi-detached, row, apartment); and second by intended markets (homeowner, rental, condo, co-op). The data is the same under the total metric but classified in these two groups. For this study, both sets were used.

Some lines of data from the census are binary variables. These are homeownership, marital status, immigration, building age, and core housing need indicators. Binary variables were handled by converting them into continuous variables. In the case of core housing variables, the census provides a line for the total number of observations separately from the subset observations. For example, core housing need is given as the total observations, number of households in core housing need, and number of households not in core housing need, as separate lines of data, where the sum of in core housing need/not in core housing need is equal to the total observations. For the other variables, the total is not given separately. In such cases, the total was calculated by adding each binary classification line of data into one. Conversion into binary variables was done by calculating each variable as a percentage of the total, using the Field Calculator tool in ArcGIS Pro. The same method was used to calculate the percentage of households classified as affordable in each census tract, under the variable name Percent_Affordable.

Inclusionary zoning was considered but not included in the spatial analysis. Inclusionary zoning will be discussed further in the Discussion section of this paper.

2.2 Classification & Regression Trees, Random Forest Modelling

R is a programming language designed for use in statistical computing. R was used in this project for the purpose of generating classification and regression trees, producing models with random forest, and Gi-* statistics hotspot mapping.

The result of the data processing in ArcGIS Pro in the prior section was a single shapefile containing 1489 rows (census tracts) and 150 columns (variables). Some variables were purely contextual, some contained spatial information necessary for programs like ArcGIS Pro to draw the feature class, some were rendered obsolete, and some were not used at all. This shapefile was exported as an Excel table and loaded into the environment within R. To begin the data analysis in R, the following R packages were used: pdp v. 0.8.3 (Greenwell 2017), randomForest v. 4.7.1.2 (Liaw and Wiener 2002), rnaturalearth v. 1.2.0 (Massicotte and South 2026), rpart v. 4.1.27 (Therneau and Atkinson 2026), rpart.plot v. 3.1.4 (Milborrow 2026), sf v. 1.1.1 (Pebesma 2018; Pebesma and R. Bivand 2023), spdep v. 1.4.2 (Bivand et al. 2013; Bivand and Wong 2018; Roger Bivand 2022; Pebesma and R. S. Bivand 2023), svglite v. 2.2.2 (Wickham et al. 2025), tidyverse v. 2.0.0 (Wickham et al. 2019).

A subset of the full dataset was created to consolidate the numerically continuous variables that will be used during the analysis:

Table 5. Comprehensive dataset of all continuous variables used in the analysis.

Variable Name	Description
Census_Tract_Name	Census Tract Name
Percent_Affordable	Percentage of homes classified as affordable
Percent_Suitable	Percentage of homes classified as suitable
Median_Income	Median Income
Avg_Household_Size	Average Household Size
Percent_Married	Percentage of people married
Percent_Immigrants	Percentage of immigrants
Percent_Recent_Immigrants	Percentage of immigrants landed between 2011-2021
Percent_Owner	Percentage of homeowners
Percent_Built_1960_or_earlier	Percentage of housing stock built in 1960 or earlier
Percent_Built_1961_1980	Percentage of housing stock built between 1961 and 1980
Percent_Built_1981_1990	Percentage of housing stock built between 1981 and 1990
Percent_Built_1991_2000	Percentage of housing stock built between 1991 and 2000
Percent_Built_2001_2005	Percentage of housing stock built between 2001 and 2005
Percent_Built_2006_2010	Percentage of housing stock built between 2006 and 2010
Percent_Built_2011_2015	Percentage of housing stock built between 2011 and 2015

Variable Name	Description
Percent_Built_2016_2021	Percentage of housing stock built between 2016 and 2021
Vacancy_Studio_Num	Vacancy Rate of studio apartments
Vacancy_1_Bedroom_Num	Vacancy Rate of 1-bedroom apartments
Vacancy_2_Bedroom_Num	Vacancy Rate of 2-bedroom apartments
Vacancy_3_Bedroom_Plus_Num	Vacancy Rate of 3-bedroom+ apartments
Vacancy_Total_Num	Vacancy Rate across all apartment types
Median_Shelter_Cost_Owned	Median shelter cost per month amongst homeowners
Median_Shelter_Cost_Rented	Median shelter cost per month amongst renters
Percent_Tenants_Subsidized	Percentage of tenants under subsidized housing
Transit_Score	Absolute count of transit stops located within a 1 km buffer of the census tract's centroid
Unemployment_Rate	Unemployment rate
Road_Density	Local/street level road density, as a proxy for walkability
Population_Density	Population density
Single_Housing_Completions	Number of single detached housing projects that reached completion in 2021
Semi_Detached_Housing_Completions	Number of semi-detached housing projects that reached completion in 2021
Row_Housing_Completions	Number of row housing projects that reached completion in 2021
Apartment_Housing_Completions	Number of apartment housing projects that reached completion in 2021
Homeowner_Housing_Completions	Number of homeowner housing projects that reached completion in 2021
Rental_Housing_Completions	Number of rental housing projects that reached completion in 2021
Condo_Housing_Completions	Number of condo housing projects that reached completion in 2021
Co_Op_Housing_Completions	Number of co-op housing projects that reached completion in 2021
All_Housing_Completions	Number of housing projects that reached completion in 2021
Percent_Failure_of_One	Percentage of households that fail at least one metric of housing suitability; affordability, overcrowding, or severe disrepair
Percent_Failure_of_Affordability	Percentage of households that fail the affordability metric for suitable housing
Percent_Failure_of_Overcrowding	Percentage of households that fail the overcrowding metric for suitable housing
Percent_Failure_of_Severe_Disrepair	Percentage of households that fail the severe disrepair metric for suitable housing

Variable Name	Description
Percent_Failure_of_All	Percentage of households that fail all three metrics for suitable housing
Percent_Acceptable_Housing	Percentage of households that pass all three metrics for suitable housing
Percent_Core_Housing_Need	Percentage of households that are deemed to be in core housing need
Percent_With_Mortgage	Percentage of homeowners with a mortgage
Percent_Owners_Core_Need	Percentage of homeowners that are deemed to be in core housing need
Percent_Tenants_Core_Need	Percentage of tenants that are deemed to be in core housing need
Median_Dwelling_Value	Median dwelling value

These variables listed are continuous, as in, numerical data that R understands as holding significance on a scale from 0 to the maximum value. In the data collection phase, categorical data was collected, and in some cases, numerical data was converted to categories to provide additional points for analysis. The issue with using categorical data as-is lies with the way R handles strings, it does not understand the significance of, for example, zoning being represented as "Residential", "Commercial", or "Mixed Use". The proper way to have R analyze such variables in conjunction with continuous ones is to convert categories into factors. This allows R to instead store individual strings like "Residential, Commercial, Mixed Use" as "1, 2, 3" integers under the hood, but recognizes that these are distinct categories and the random forest model will not apply splits such as "Zoning > 2.5", which is illogical given the input. The following categorical variables were converted as factors in R and appended to the subset containing the continuous variables:

Table 6. Categorical variables converted as factor within R and appended to the primary dataset.

Variable Name	Description
Zoning	The largest zoning designation within a census tract
Vacancy_Studio_Cat	Vacancy Rate of studio apartments
Vacancy_1_Bedroom_Cat	Vacancy Rate of 1-bedroom apartments
Vacancy_2_Bedroom_Cat	Vacancy Rate of 2-bedroom apartments
Vacancy_3_Bedroom_Plus_Cat	Vacancy Rate of 3-bedroom+ apartments
Vacancy_Total_Cat	Vacancy Rate across all apartment types
Pop_Den_Category	Population density
Walkability	Classifications of local/street level road density

In total, the subset contained 57 variables that were used during the analysis. Regression trees were produced using the rpart package, and different variables were swapped in and out to compare the results of each tree. The randomForest package was then used to generate models of 1000 regression trees, with each iteration of model being improved upon given the information learned from the last. Improvements were considered based on the root mean square error of each variable in the model, and statistics such as the percentage of variables explained and number of trees required to produce the result. The randomForest package has a built-in tuner, tuneRF, which was used once a satisfactory random forest model was produced through numerous trials, allowing for an even more statistically significant result. At each split in a Random Forest model, only a randomly selected subset of the predictor variables is considered rather than the full set of variables. In R, the default number of variables evaluated at each split (the mtry parameter) is determined using a rule of thumb based on the number of predictors in the dataset. However, this default value is not necessarily optimal, so the tuneRF function was used to identify the value that minimized prediction error for this study.

```
census_rf_tuned_data <- census_subset %>%
  select(all_of(c("Percent_Affordable", final_features))) %>%
  na.omit()

X_features <- census_rf_tuned_data %>% select(-Percent_Affordable)
Y_target   <- census_rf_tuned_data$Percent_Affordable

svg("optimal_mtry_tune_plot.svg", width = 6.5, height = 4.5)

set.seed(1)
optimal_mtry_results <- tuneRF(
  x = X_features,
  y = Y_target,
  mtryStart = 9,
  ntreeTry = 1000,
  stepFactor = 1.5,
  improve = 0.001,
  trace = TRUE,
```

```
  plot = TRUE
)
```

Completion of the final produced random forest model was accompanied by a global directional effects matrix of each variable to represent its predictive power for housing affordability, and its relationship, positive or negative.

2.3 G_i* Statistics

Getis-Ord G_i* statistics were calculated from the results of the final random forest model, with spatial coordinates determined from the shapefile used in ArcGIS Pro preprocessing. This required linking the data in R, since the random forest model has no geographic context and requires a unique identification field from the original dataset to be given context. In this case, the field "CTUID", a unique identifier field for each census tract, was used to give spatial context to the results of each row in the random forest analysis. The purpose of computing G_i* statistics was to identify clusters of census tracts that the random forest model predicted incorrectly, to identify statistically significant spatial clusters where the rate of affordability is unexplained by the variables used in the analysis.

Three figures were produced; a map showing the hot/cold spatial clusters with varying levels of confidence; a bell curve showing the density of model residuals, and; a volcano plot to identify the distribution of z-scores for all plot points.

2.4 Data Gaps

There are a handful of data gaps present in this study. A small number of census tracts contain missing information in crucial fields, such as shelter-cost-to-income ratio. Of course, such census tracts could not be analyzed as this is the metric upon which the independent variable, percentage of affordable housing, is based upon. There was an attempt to consider the line "Dwelling provided by the local government, First Nation or Indian band" and its relevant data from the census, but the only census tract that has this characteristic is the Chippewas of Georgina Island First Nation, which is missing critical data points like STIR. Random Forest models are

unable to consider any rows which have missing data, as such, approximately ~200 of the ~1500 census tracts data was collected for were not included in the results.

3. Results

The independent variable for random forest modelling is “Percent_Affordable”. This is a field which was calculated using the census data from the shelter cost to income ratio (STIR) category. The number of households identified as spending less than 30% of their income on shelter costs per month was divided by the total number of households surveyed to calculate the percentage of affordable housing for each census tract.

Before running any random forest models, the global randomization seed within R was set to 1. This is done purposefully to allow for reproducibility with the model’s results for future research and does not impact the viability of the results by making the model’s randomness inherently less reliable. With random forest models, a statistical output is generated:

- No. of variables tried at each split (mtry)
    - This means that at every single branch node split across all trees generated (in this case, 1000 trees were generated for each model), the algorithm randomly ignored most of the predictors and forced the tree to choose the best split from a randomized subset of exactly X variables. This typically ranged from 9-11 in the mathematical formula used by R to try guessing the most optimal value to use, however it was optimized through tuning later in the analysis to produce a tighter model.
- Mean of squared residuals
    - This is the Mean Squared Error (MSE) of the model. It is calculated by taking the difference between the actual affordability score in each census tract and the model’s prediction, squaring the result, and averaged across the entire dataset. The square root of this average can be taken to get the Root Mean Squared Error (RMSE), giving a percentage of how off the model’s predictions are on average.
- % Var explained
    - This is the model’s overall grade, the Out-of-Bag R^2 score. It represents the percentage of independent socio-demographic features in the dataset that were successfully captured of the total housing affordability values across the region.
- which.min(RFMODEL$mse)
    - This value represents the tree in which its lowest error rate hit. It means that every tree generated after this value only served to remove mathematical fluctuations.

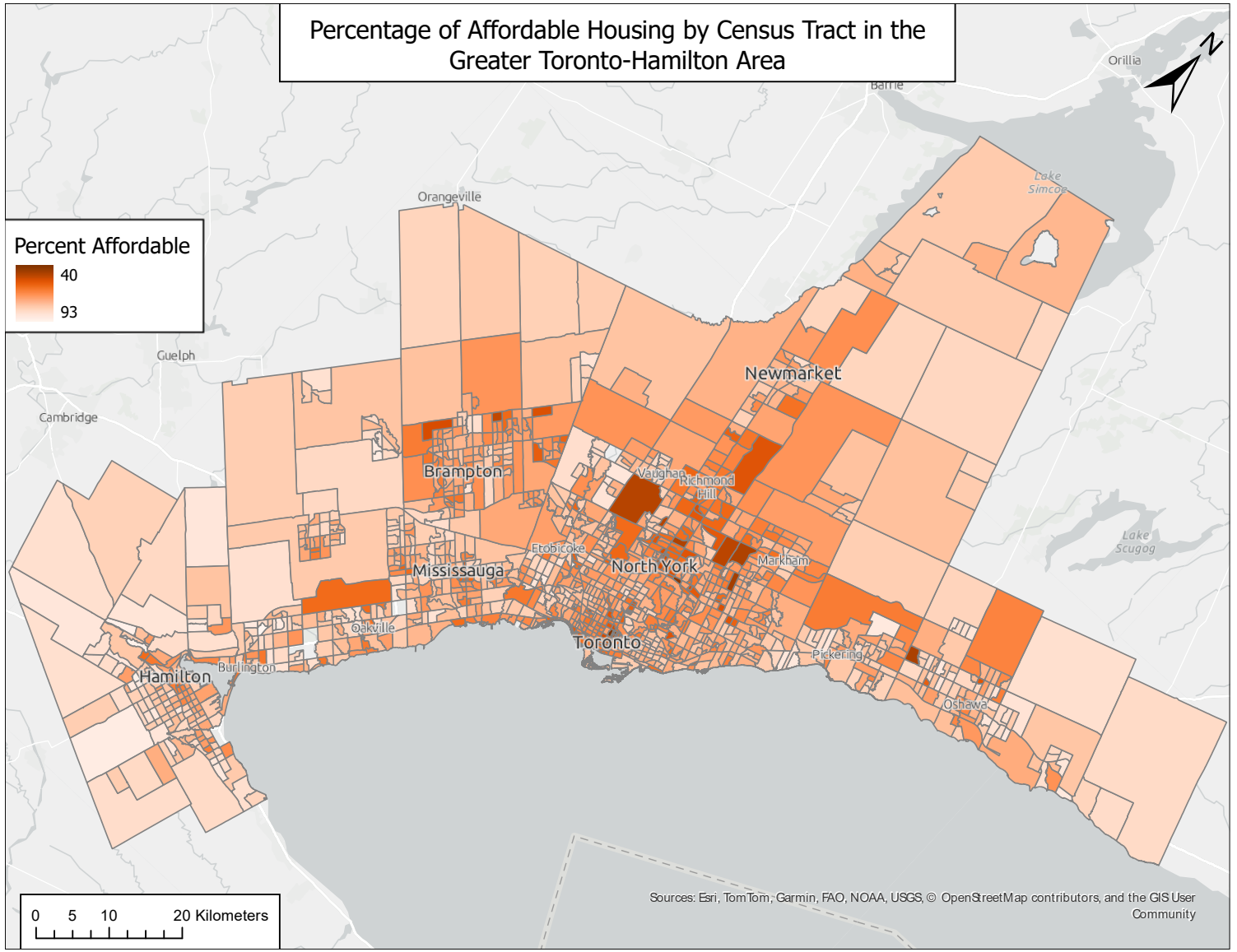


Figure 6. Percentage of Affordable Housing by Census Tract in the GTHA.

3.1 Decision Trees & Random Forest Models

The first decision tree contained all the variables.

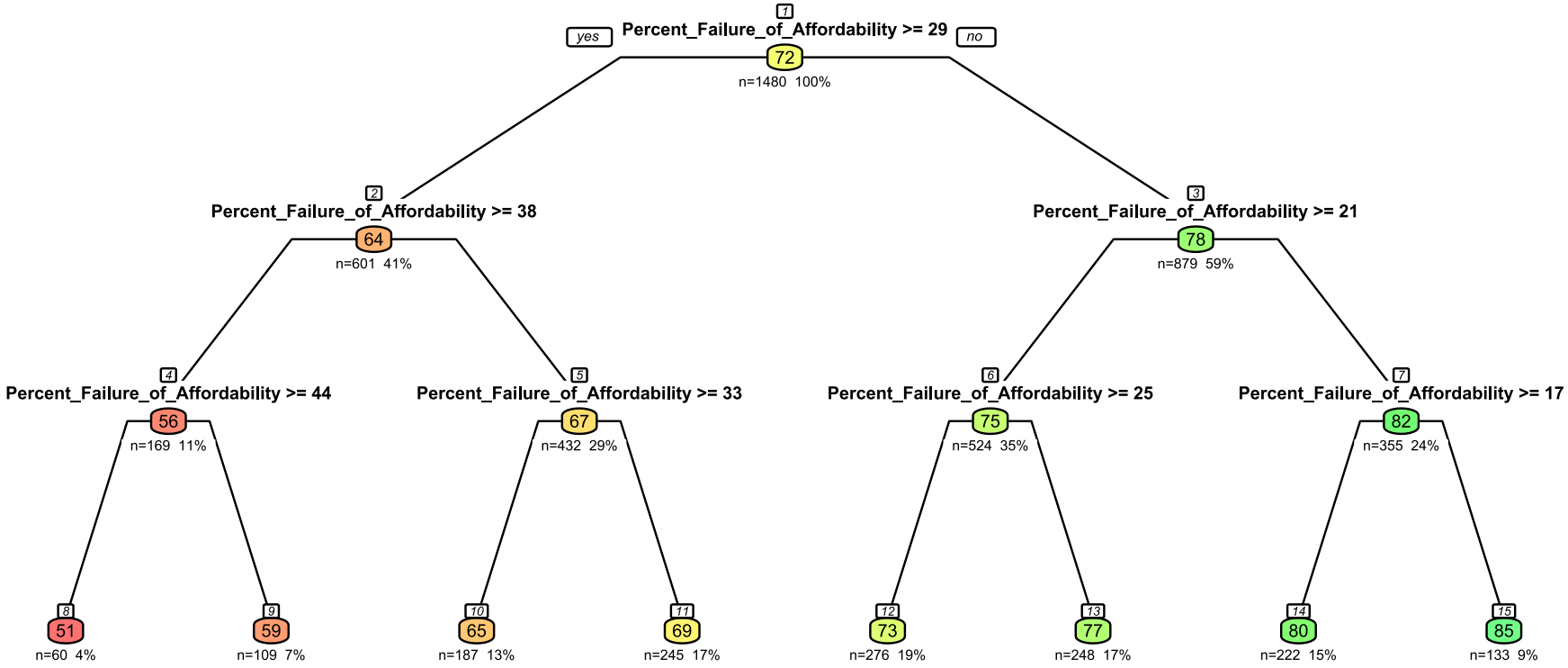


Figure 7. Decision Tree with all variables.

```
Variables actually used in tree construction:
[1] Percent_Failure_of_Affordability


Root node error: 105090/1480 = 71.007


n=1480 (9 observations deleted due to missingness)
```

Table 7. Decision tree results using all variables.

	CP	nsplit	rel error	xerror	xstd
1	0.645115	0	1.000000	1.000328	0.0388503
2	0.135412	1	0.354885	0.356200	0.0160473
3	0.108119	2	0.219473	0.220884	0.0086299
4	0.021728	3	0.111354	0.112696	0.0056428
5	0.020238	4	0.089626	0.092634	0.0043363
6	0.018083	5	0.069387	0.070036	0.0033381
7	0.017357	6	0.051305	0.058142	0.0032082
8	0.010000	7	0.033948	0.035408	0.0023973

In this first decision tree, all variables were thrown in together to see what would happen. One of these variables includes “Percent Failure of Affordability”, which is one of the measures used to determine core housing need. Of course, this variable being included is counterintuitive to the goal of analysing predictors for “Percent Affordable”. The resulting tree only making decisions to split based on “Percent Failure of Affordability” is a good sign for data integrity, but this variable cannot be used for further analysis.

The next tree removes affordability issues from the tree to see which variables are influencing the independent, “Percent Affordable”.

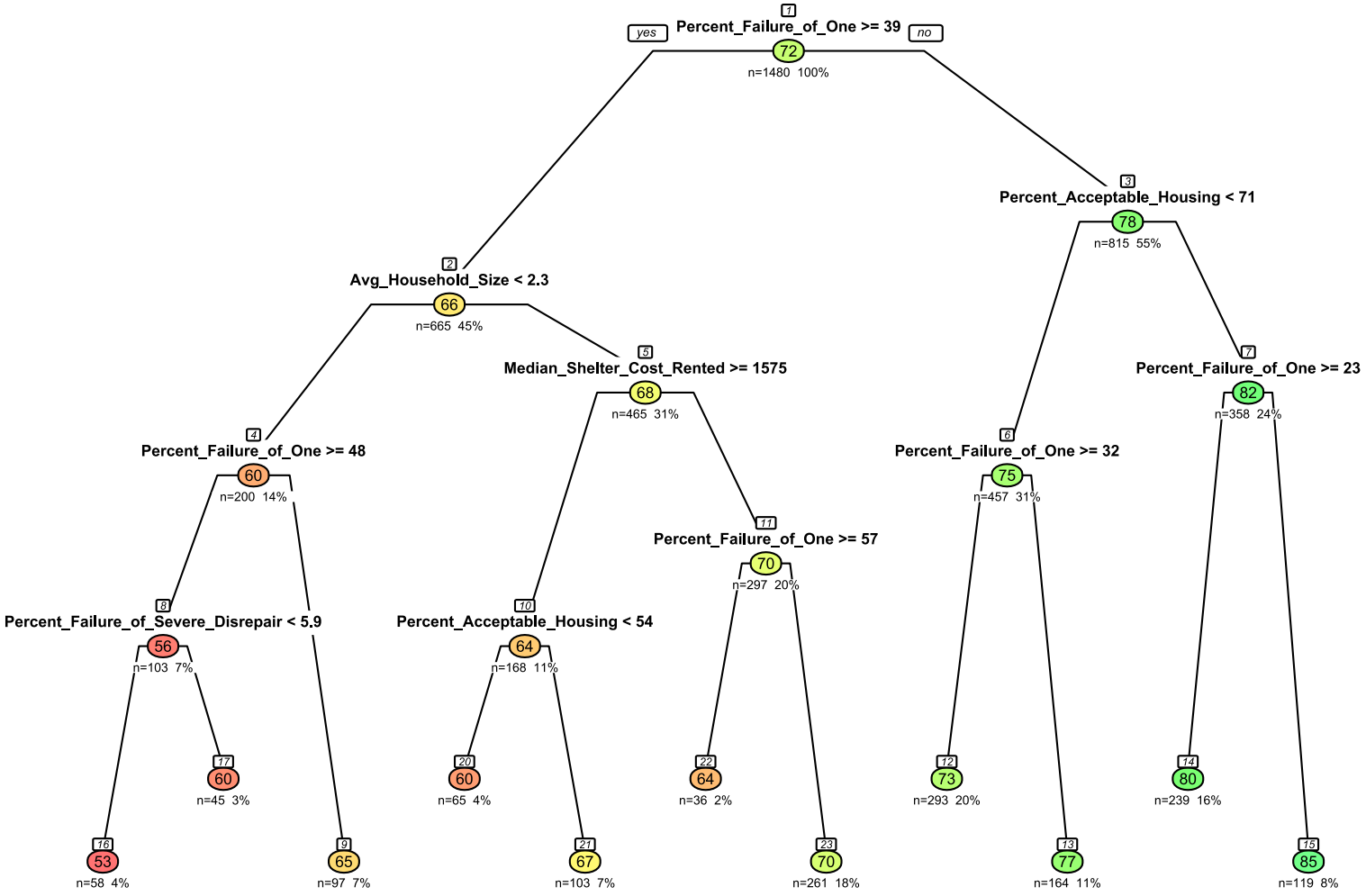


Figure 8. Decision Tree without affordability variables.

```
Variables actually used in tree construction:

[1] Avg_Household_Size                  Median_Shelter_Cost_Rented
Percent_Acceptable_Housing

[4] Percent_Failure_of_One              Percent_Failure_of_Severe_Disrepair


Root node error: 105090/1480 = 71.007


n=1480 (9 observations deleted due to missingness)
```

Table 8. Decision Tree results without affordability variables.

	CP	nsplit	rel error	xerror	xstd
1	0.521293	0	1.00000	1.00280	0.0389854
2	0.096441	1	0.47871	0.48348	0.0203200
3	0.072079	2	0.38227	0.38928	0.0196990
4	0.036677	3	0.31019	0.33212	0.0151142
5	0.028957	4	0.27351	0.30312	0.0139452
6	0.018042	5	0.24455	0.26170	0.0116579
7	0.016797	6	0.22651	0.25189	0.0114100
8	0.012850	7	0.20971	0.23964	0.0112632
9	0.012479	8	0.19686	0.22314	0.0102216
10	0.011656	9	0.18438	0.21456	0.0101171
11	0.010000	10	0.17273	0.20374	0.0094601

In this tree, “Percent Failure of One” and “Acceptable Housing” are prevalent. Both variables are subsets of the definition for core housing need, which includes affordability as one of its measures. These two variables, along with “Failure of all Three” have direct correlation with housing affordability, but the relationship is not 1:1 like it is with “Percent Failure of Affordability”, which allows for other variables to have some influence on the decision tree. Notably, the failure of all three housing indicators does not show up at all.

The next tree removes “Percent Failure of One” core housing indicator and “Percent Acceptable Housing” from the tree to see if more or different variables will influence the independent.

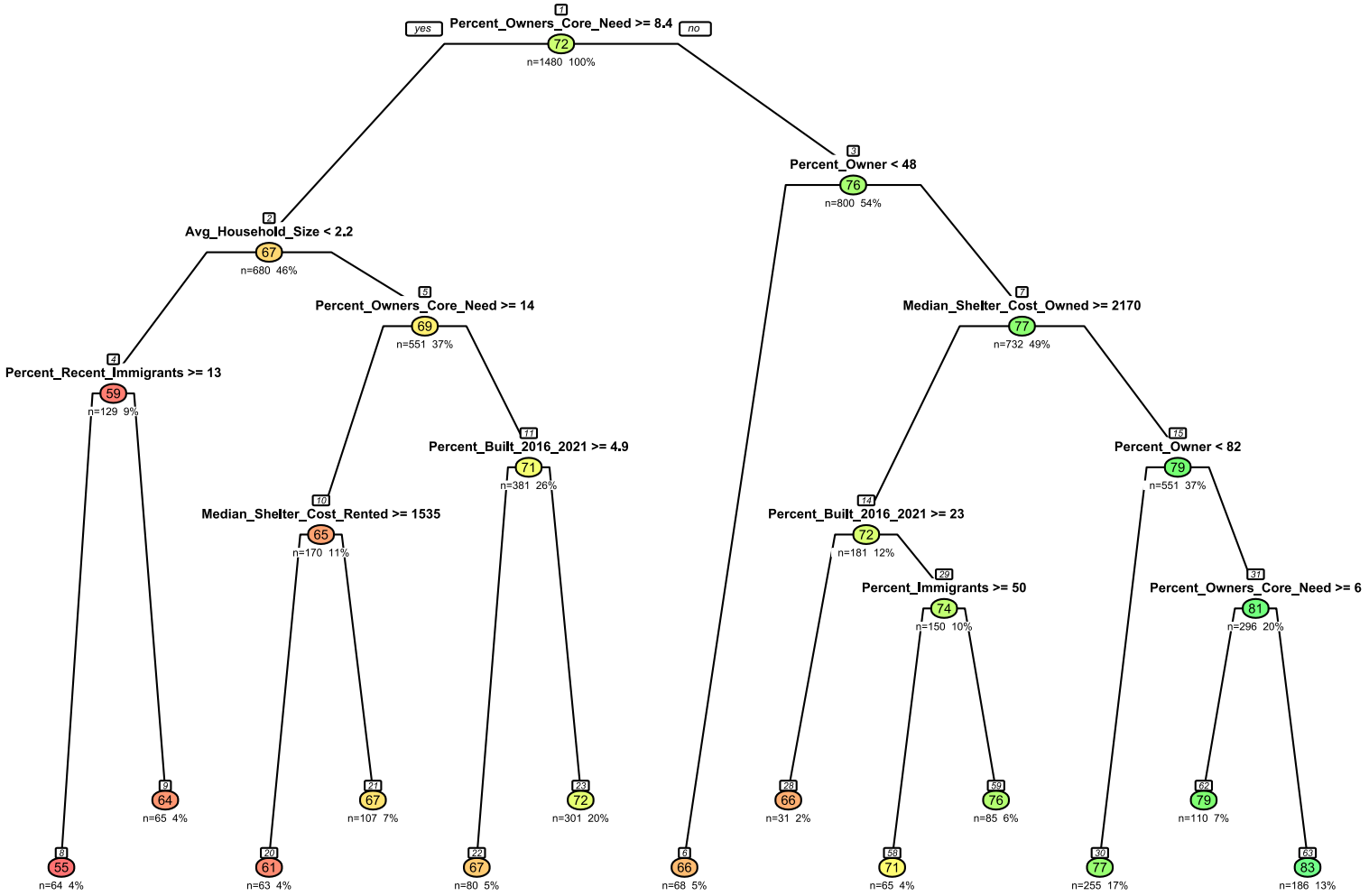


Figure 9. Decision Tree without core housing indicator variables.

```
Variables actually used in tree construction:

[1] Avg_Household_Size         Median_Shelter_Cost_Owned  Median_Shelter_Cost_Rented
Percent_Built_2016_2021

[5] Percent_Immigrants         Percent_Owner             Percent_Owners_Core_Need
Percent_Recent_Immigrants


Root node error: 105090/1480 = 71.007


n=1480 (9 observations deleted due to missingness)
```

Table 9. Decision Tree results without core housing indicator variables.

	CP	nsplit	rel error	xerror	xstd
1	0.291839	0	1.00000	1.00130	0.038932
2	0.098792	1	0.70816	0.72587	0.030120
3	0.074666	2	0.60937	0.63263	0.026172
4	0.061014	3	0.53470	0.57955	0.024874
5	0.040145	4	0.47369	0.52848	0.022108
6	0.027465	5	0.43354	0.49556	0.021046
7	0.023140	6	0.40608	0.46369	0.019799
8	0.016038	7	0.38294	0.45050	0.019624
9	0.015930	8	0.36690	0.44033	0.018960
10	0.014363	9	0.35097	0.43657	0.018868
11	0.011089	10	0.33661	0.41326	0.017545
12	0.010219	11	0.32552	0.39711	0.016735
13	0.010000	12	0.31530	0.38982	0.016408

In this tree, “Percent Owners Core Need” is dominating the decision tree. Notably, the variable “Percent Tenants Core Need” went unused in the decision algorithm, despite “Percent Owner” holding significance in the model, and "Percent Failure of All” suitable housing indicators is also not a consideration. This is most likely due to a lack of sample size for the number of households that severely lack adequate housing by failing all three measures for core housing need.

The next tree will remove “Percent Owners Core Need”, since it is a near direct correlation of affordability and will thus reduce bias towards measures that inherently consider affordability directly. It will, however, leave in “Percent Tenants Core Need” since it has not shown up in any of the decision trees so far.

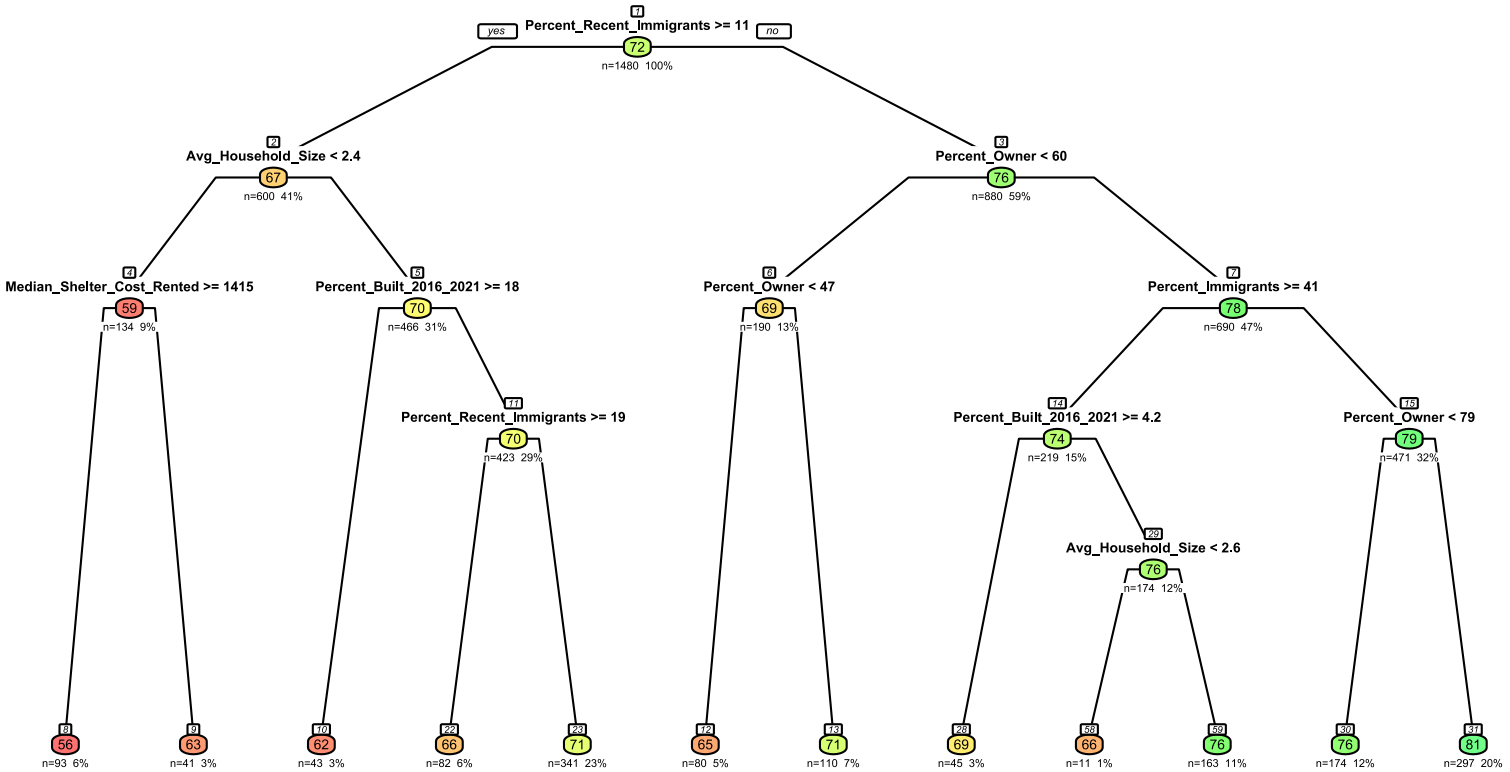


Figure 10. Decision Tree without homeowners in core need.

```
Variables actually used in tree construction:

[1] Avg_Household_Size         Median_Shelter_Cost_Rented Percent_Built_2016_2021
Percent_Immigrants

[5] Percent_Owner              Percent_Recent_Immigrants


Root node error: 105090/1480 = 71.007


n=1480 (9 observations deleted due to missingness)
```

Table 10. Decision Tree results without homeowners in core need.

	CP	nsplit	rel error	xerror	xstd
1	0.253623	0	1.00000	1.00116	0.038934
2	0.121391	1	0.74638	0.79100	0.031921
3	0.117157	2	0.62499	0.69885	0.030161
4	0.034803	3	0.50783	0.53219	0.023699
5	0.024052	4	0.47303	0.50895	0.022678
6	0.022882	5	0.44897	0.49612	0.021648
7	0.018519	6	0.42609	0.48157	0.021445
8	0.017129	7	0.40757	0.46153	0.020644
9	0.015222	8	0.39044	0.45450	0.020627
10	0.013286	9	0.37522	0.45115	0.020499
11	0.010898	10	0.36194	0.44252	0.020610
12	0.010000	11	0.35104	0.42966	0.020038

This tree is a good baseline to use, since most of the chosen variables do not necessarily have a direct correlation with affordable housing and are not unfairly influencing the model's choices. These variables are "Percent Recent Immigrants", "Average Household Size", "Percent Owner", and "Percent Built 2016 to 2021", a subset of building age. "Median Shelter Cost Rented" has a presumably larger direct influence on affordable housing, since it is part of the calculation for shelter-cost-to-income ratio, which is being used as the indicator for percentage of affordable housing.

Next, random forest models were built using this tree as a baseline. The following figures depict variable predictor importance rankings for each random forest model. The greater the value of the horizontal axis, the increase in root mean square error, the greater the predictive power of the variable on the model (whether it has a positive or negative influence on affordability).

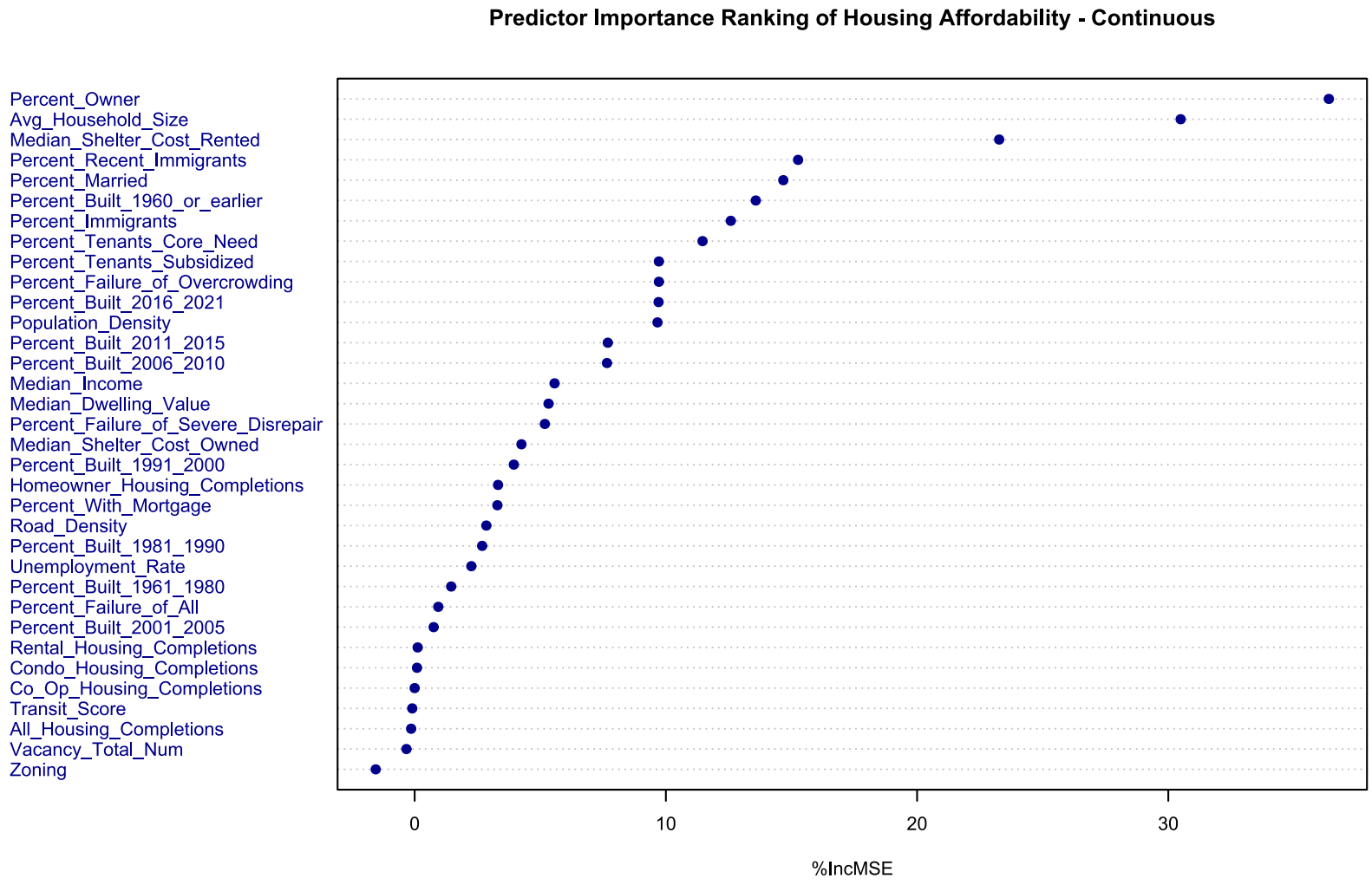


Figure 11. Variable predictor importance ranking from the first random forest model.

```
               Type of random forest: regression
                     Number of trees: 1000
No. of variables tried at each split: 11


          Mean of squared residuals: 17.04312
                    % Var explained: 65.43
> which.min(rf_continuous_tree$mse)
[1] 669
> varImpPlot(rf_continuous_tree, type = 1, main = "Top Drivers of Housing
Affordability")
> # How many rows were dropped, how many rows were  used?
> num_omitted <- length(rf_continuous_tree$na.action)
> print(paste("Rows omitted:", num_omitted))
[1] "Rows omitted: 1265"
>
> num_kept <- length(rf_continuous_tree$y)
> print(paste("Rows kept/used:", num_kept))
[1] "Rows kept/used: 224"
```

This first random forest model was created using a majority of “continuous” variables rather than categorical counterparts. Examples include “Vacancy Total Num”, “Road Density”, and “Population Density”. The main issue with this output is the number of rows omitted, 1265, compared to rows kept/used, 224. The reason this happened is because random forest models can only use rows in which no values are left blank, or in technical terms, null. There are cases where some census tracts were lacking critical data related to this study, such as the data regarding STIR, in which cases data omission is unavoidable. In this particular case, the enormous data loss is due to the lack of data available for vacancy rate at a census tract level, and how that data is represented as a continuous variable. When vacancy rate data is missing for a particular census tract, it is defined as null, which causes the entire row to be thrown out by the random forest model. To amend this problem, vacancy rate was categorized as a factor variable, distributed into categories “N/A, Zero, Low, Medium, High”, as outlined in the Methods.

The next iteration of the random forest model attempts to amend this not only with vacancy rate, but all other categorical possibilities as a comparison point.

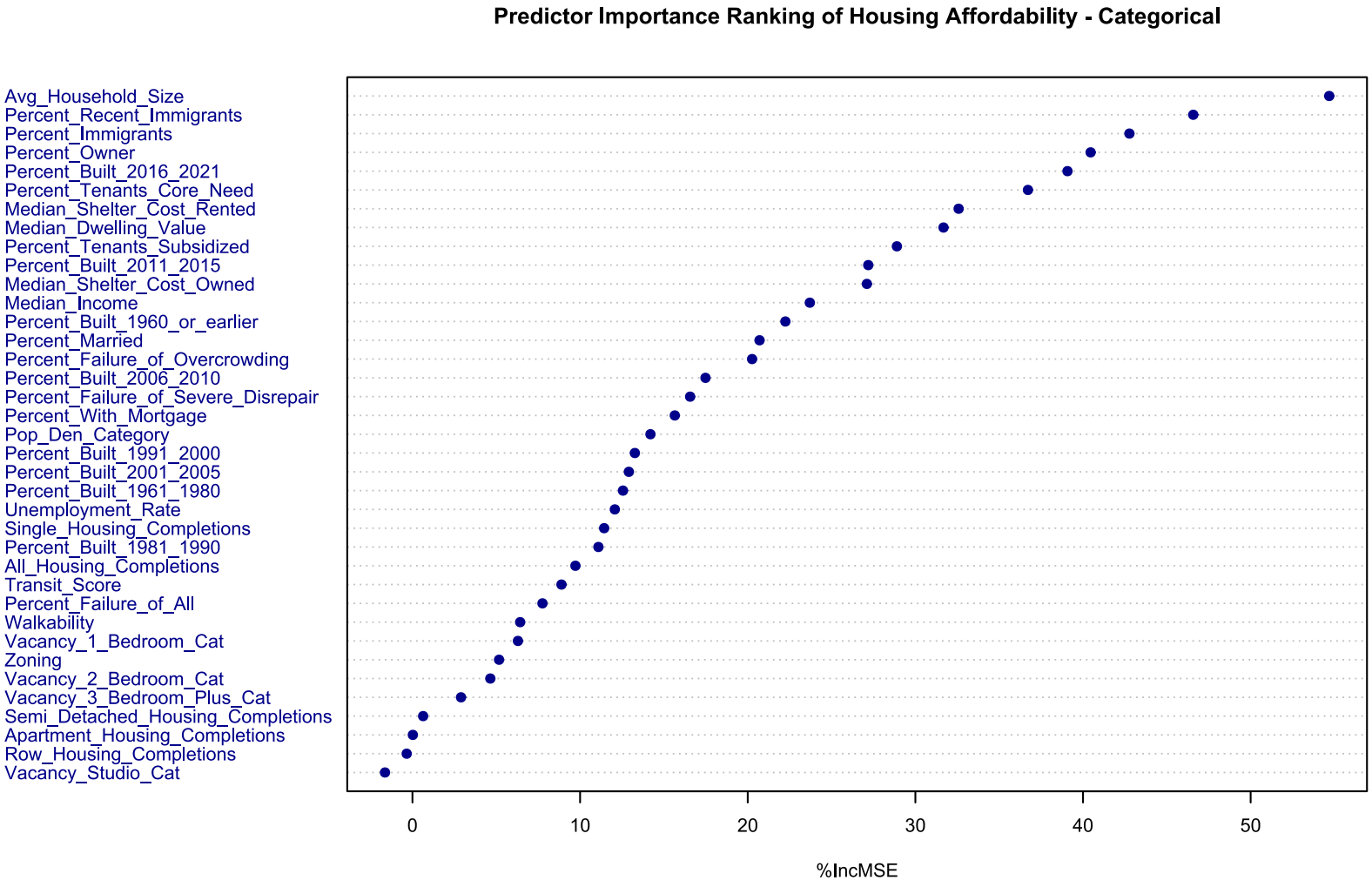


Figure 12. Variable predictor importance ranking from the categorical-focused random forest model.

```
               Type of random forest: regression
                     Number of trees: 1000
No. of variables tried at each split: 12


          Mean of squared residuals: 13.05482
                    % Var explained: 80
> which.min(rf_categorical_tree$mse)
[1] 974
> varImpPlot(rf_categorical_tree, type = 1, main = "Top Drivers of Housing
Affordability")
> # How many rows were dropped, how many rows were  used?
> num_omitted <- length(rf_categorical_tree$na.action)
> print(paste("Rows omitted:", num_omitted))
[1] "Rows omitted: 179"
>
> num_kept <- length(rf_categorical_tree$y)
> print(paste("Rows kept/used:", num_kept))
[1] "Rows kept/used: 1310"
```

This model is far more useful in its predictive power due to the usable rows growing to 1310. The statistics produced by the model are now meaningful for comparison:

- Variables tried at each split (mtry): 12
- Mean of squared residuals: 13.05482
- Variables explained: 80%
- Minimum number of trees required: 974

The next model removes “All Housing Completions” to prevent double counting the same data, and swaps the categories for housing tenure for housing completions from “Single Housing Completions”, “Semi Detached Housing Completions”, “Row Housing Completions”, and “Apartment Housing Completions” to “Homeowner Housing Completions”, “Rental Housing Completions”, “Condo Housing Completions”, and “Co-op Housing Completions”. It also swaps the different apartment sizes for vacancy rate in favour of “Vacancy Total Cat”. This was to give different perspectives on changing categorizations of variables.

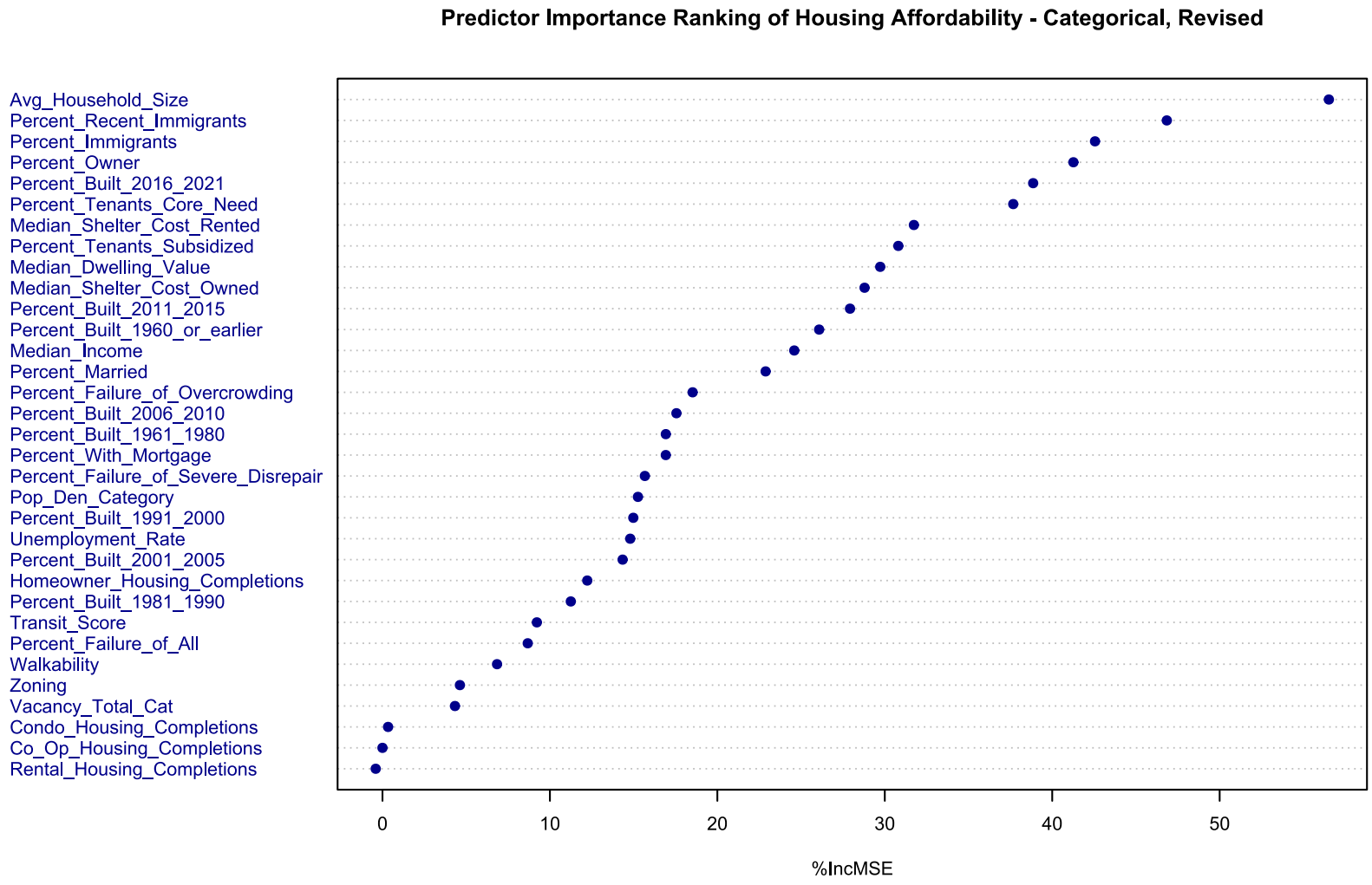


Figure 13. Variable predictor importance ranking from the random forest model with consolidation of vacancy rate.

```
               Type of random forest: regression
                     Number of trees: 1000
No. of variables tried at each split: 11


          Mean of squared residuals: 12.95771
                    % Var explained: 80.15
> which.min(rf_categorical_tree_revised$mse)
[1] 626
> varImpPlot(rf_categorical_tree_revised, type = 1, main = "Top Drivers of Housing
Affordability")
> # How many rows were dropped, how many rows were  used?
> num_omitted <- length(rf_categorical_tree_revised$na.action)
> print(paste("Rows omitted:", num_omitted))
[1] "Rows omitted: 179"
>
> num_kept <- length(rf_categorical_tree_revised$y)
> print(paste("Rows kept/used:", num_kept))
[1] "Rows kept/used: 1310"
```

- Variables tried at each split (mtry): 11
- Mean of squared residuals: 12.95771
- Variables explained: 80.15%
- Minimum number of trees required: 626

Based on Figure 13, “Homeowner Housing Completions” shows some predictive power, but “Rental Housing Completions”, “Condo Housing Completions”, and “Co op Housing Completions” are either near zero or negative for increases in mean squared error for the model. These three variables can be discarded as noise and shows that separating “Total Housing Completions” into these categories has allowed the model to isolate housing completions with the intended target market as homeowners to be the only type of completions with predictive power.

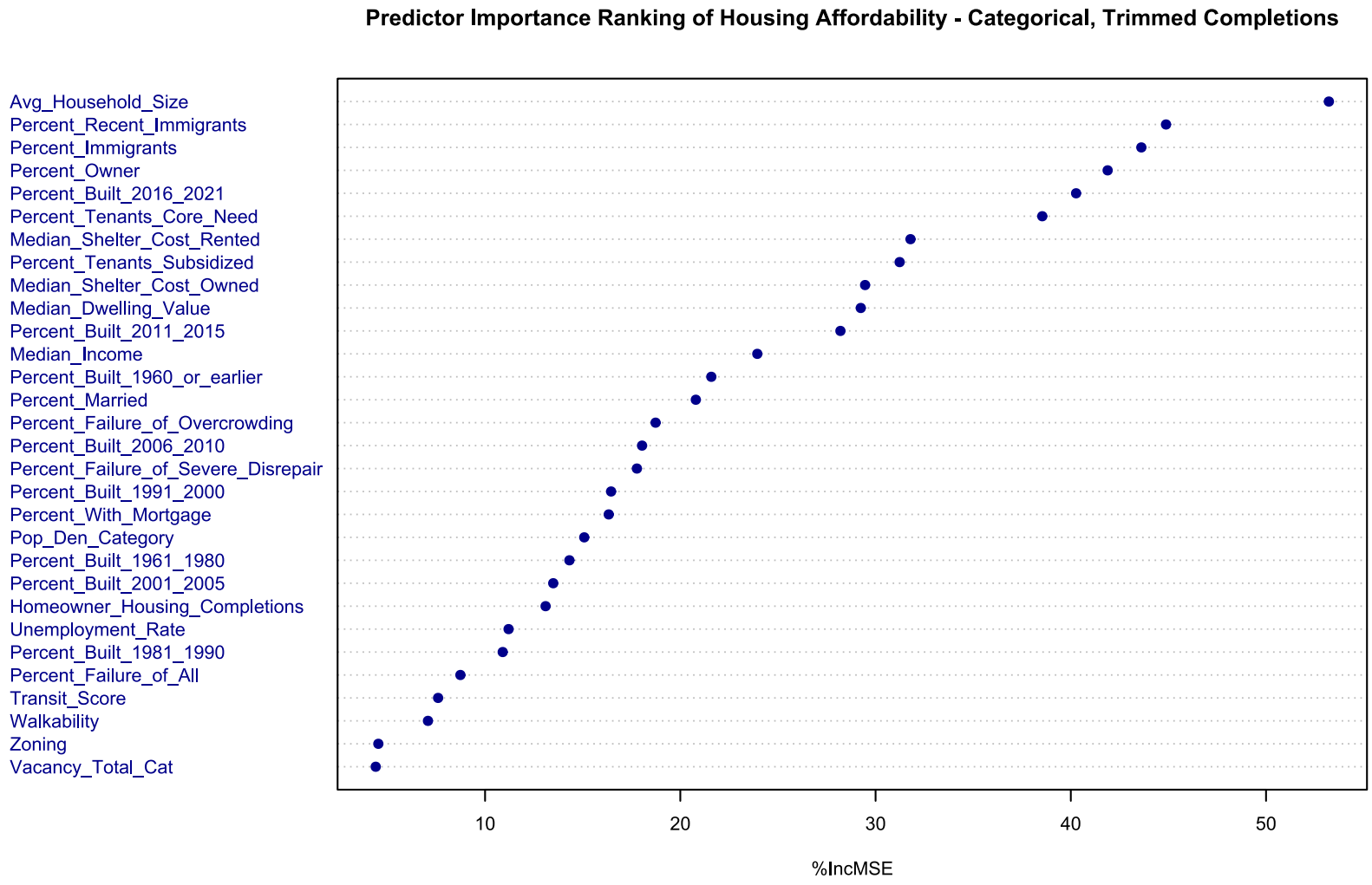


Figure 14. Variable predictor importance ranking from the random forest model with trimmed housing completion variables.

```
               Type of random forest: regression
                     Number of trees: 1000
No. of variables tried at each split: 10


          Mean of squared residuals: 13.00735
                    % Var explained: 80.08
> which.min(rf_categorical_tree_trimmed_completions$mse)
[1] 1000
> varImpPlot(rf_categorical_tree_trimmed_completions, type = 1, main = "Top Drivers of
Housing Affordability")
> # How many rows were dropped, how many rows were  used?
> num_omitted <- length(rf_categorical_tree_trimmed_completions$na.action)
> print(paste("Rows omitted:", num_omitted))
[1] "Rows omitted: 179"
>
> num_kept <- length(rf_categorical_tree_trimmed_completions$y)
> print(paste("Rows kept/used:", num_kept))
[1] "Rows kept/used: 1310"
```

- Variables tried at each split (mtry): 10
- Mean of squared residuals: 13.00735
- Variables explained: 80.08%
- Minimum number of trees required: 1000

The next model will replace “Walkability” with “Road Density” to compare the predictive power between the two, as they are the same variable represented differently by continuous/categorical values.

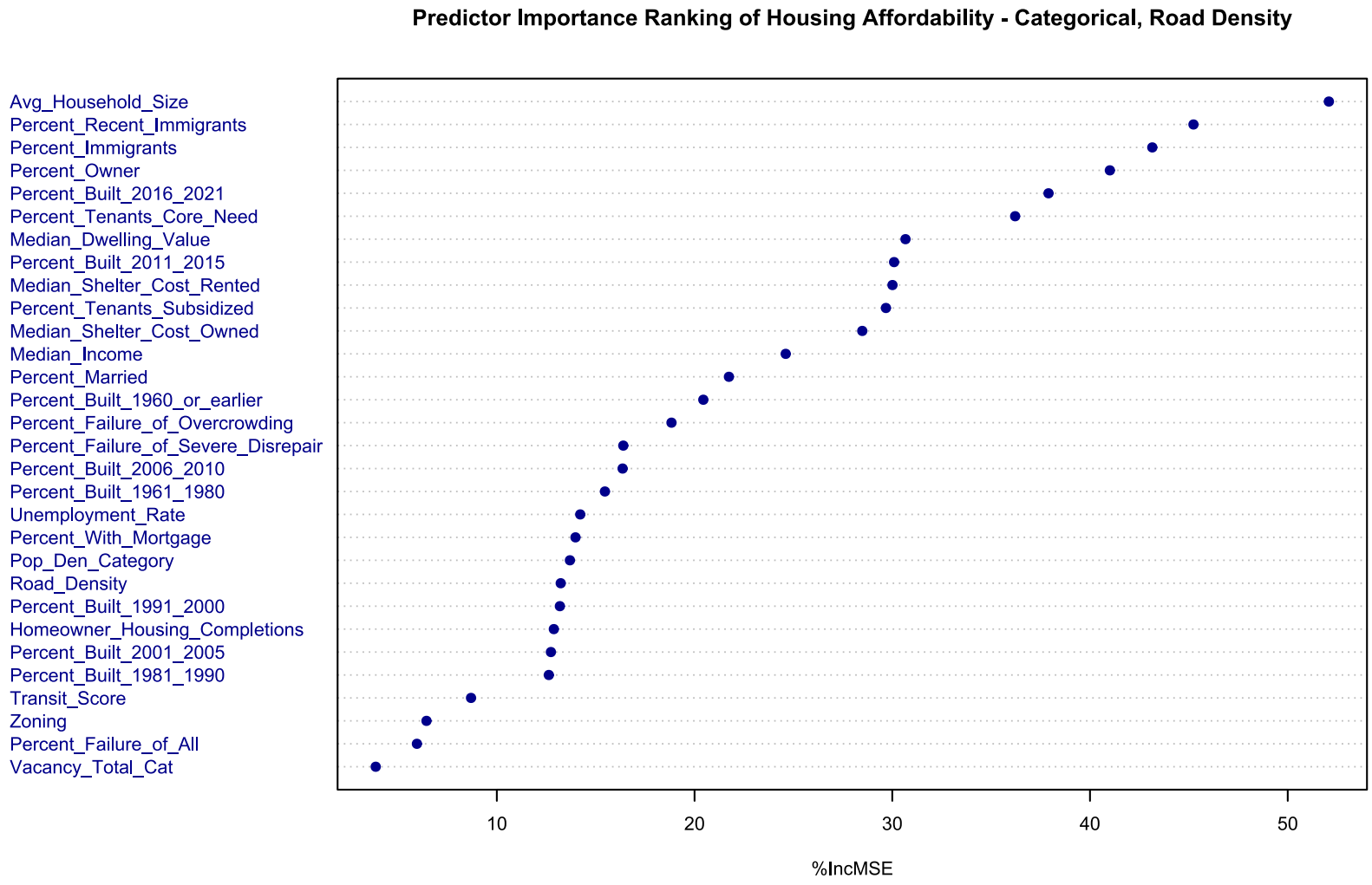


Figure 15. Variable predictor importance ranking from the random forest model with road density instead of walkability.

```
               Type of random forest: regression
                     Number of trees: 1000
No. of variables tried at each split: 10


          Mean of squared residuals: 12.94541
                    % Var explained: 80.08
> which.min(rf_categorical_tree_road_density$mse)
[1] 552
> varImpPlot(rf_categorical_tree_road_density, type = 1, main = "Top Drivers of
Housing Affordability")
> # How many rows were dropped, how many rows were  used?
> num_omitted <- length(rf_categorical_tree_road_density$na.action)
> print(paste("Rows omitted:", num_omitted))
[1] "Rows omitted: 181"
>
> num_kept <- length(rf_categorical_tree_road_density$y)
> print(paste("Rows kept/used:", num_kept))
[1] "Rows kept/used: 1308"
```

- Variables tried at each split (mtry): 10
- Mean of squared residuals: 12.94541
- Variables explained: 80.08%
- Minimum number of trees required: 552

The next model will do the same regarding categorical variables into continuous, except this time replace “Pop Den Cat” with “Population Density”.

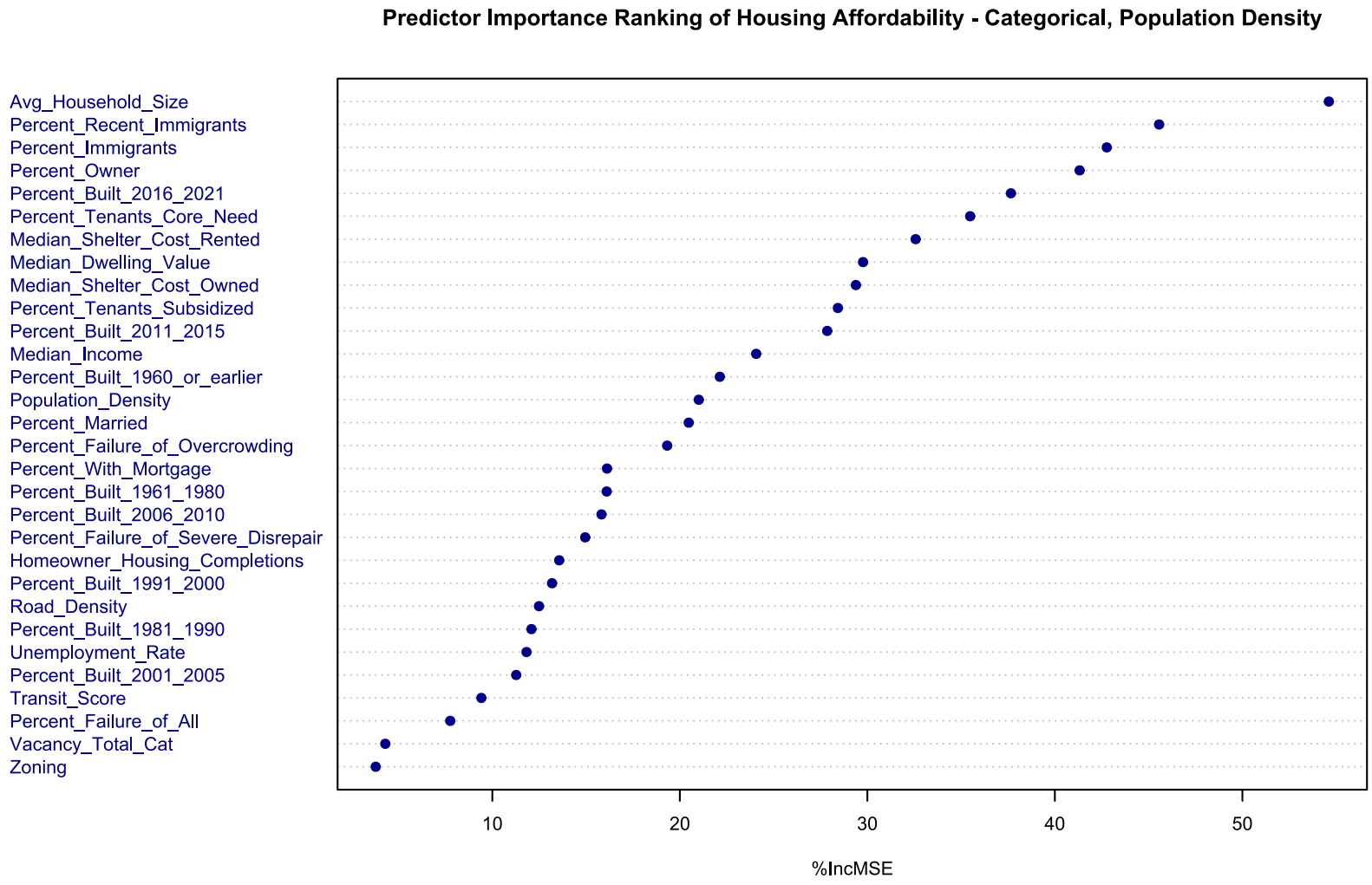


Figure 16. Variable predictor importance ranking from the random forest model with population density as a continuous variable.

```
               Type of random forest: regression
                     Number of trees: 1000
No. of variables tried at each split: 10


          Mean of squared residuals: 13.02517
                    % Var explained: 79.96
> which.min(rf_categorical_tree_pop_den$mse)
[1] 987
> varImpPlot(rf_categorical_tree_pop_den, type = 1, main = "Top Drivers of Housing
Affordability")
> # How many rows were dropped, how many rows were  used?
> num_omitted <- length(rf_categorical_tree_pop_den$na.action)
> print(paste("Rows omitted:", num_omitted))
[1] "Rows omitted: 181"
>
> num_kept <- length(rf_categorical_tree_pop_den$y)
> print(paste("Rows kept/used:", num_kept))
[1] "Rows kept/used: 1308"
```

- Variables tried at each split (mtry): 10
- Mean of squared residuals: 13.02517
- Variables explained: 79.96%
- Minimum number of trees required: 987

It is worth considering the impacts of “Percent Tenants Core Need” and “Percent Failure of All” on unfairly influencing the decisions made by the model. The percentage of owners in core need was removed from the model due to inherently considering affordability as part of its characteristics as a variable, and percentage failing one core housing metric was also removed for the same reason. The model may tend to “cheat” if given variables that give away information about general affordability within a census tract that other variables can not necessarily explain. Although these are interesting variables to see plotted, ultimately it is counterintuitive to include these when trying to model the predictive power of variables for affordability outcomes. Thus, these two variables were removed from the next model.

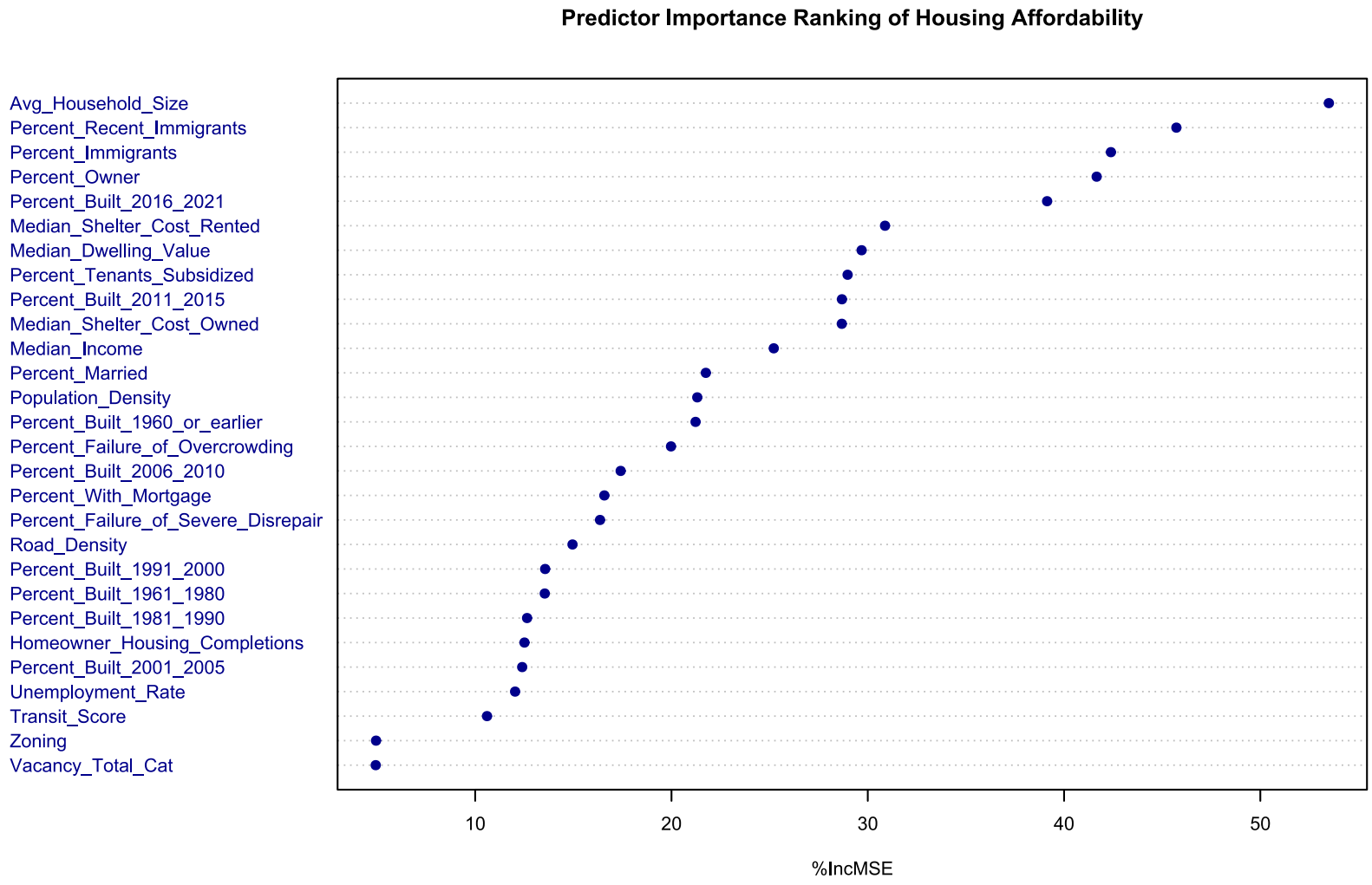


Figure 17. Variable predictor importance ranking from the random forest model with removing the remainder of core housing need variables.

```
               Type of random forest: regression
                     Number of trees: 1000
No. of variables tried at each split: 9


          Mean of squared residuals: 13.72762
                    % Var explained: 78.88
> which.min(rf_final_model$mse)
[1] 982
> varImpPlot(rf_final_model, type = 1, main = "Top Drivers of Housing Affordability")
> # How many rows were dropped, how many rows were  used?
> num_omitted <- length(rf_final_model$na.action)
> print(paste("Rows omitted:", num_omitted))
[1] "Rows omitted: 181"
>
> num_kept <- length(rf_final_model$y)
> print(paste("Rows kept/used:", num_kept))
[1] "Rows kept/used: 1308"
```

- Variables tried at each split (mtry): 9
- Mean of squared residuals: 13.72762
- Variables explained: 78.88%
- Minimum number of trees required: 982

The next model will be tuned using the randomForest R package built in tuner, tuneRF. The model attempts to pick the number of variables to try at each split based on a mathematical formula, although this assumption may not always give the best error approximation.

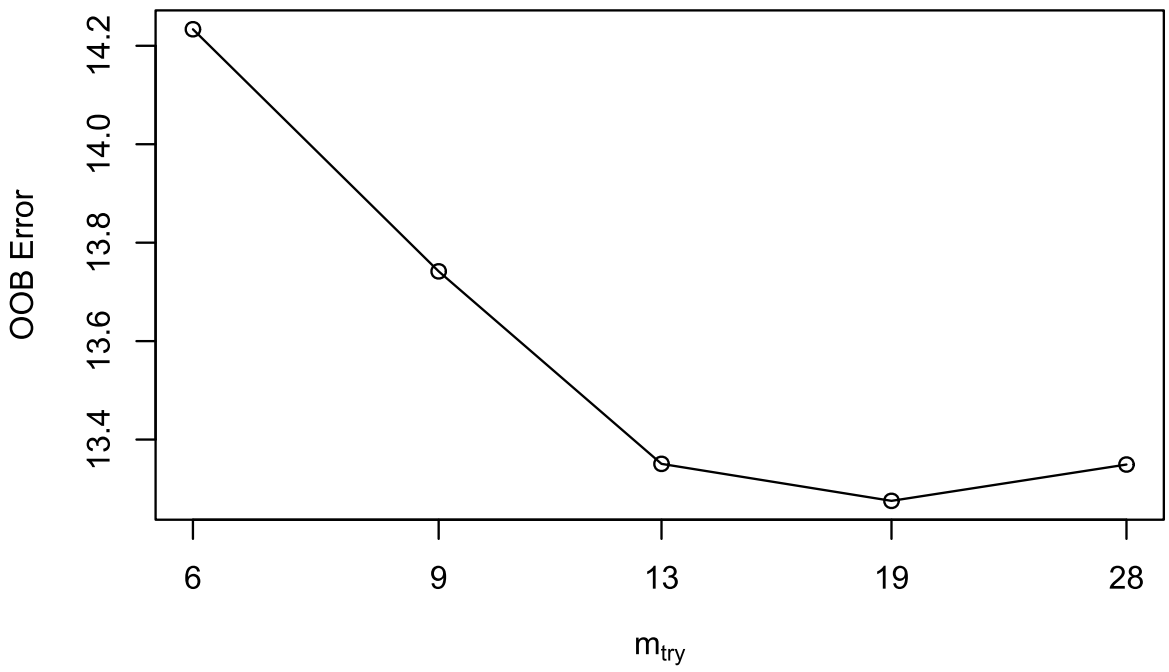


Figure 18. Results of tuneRF Function to Demonstrate Optimal Value of mtry.

The tuneRF function searches for the optimal value of mtry, the number of variables to try at each given split, with respect to the Out-of-Bag (OOB) error estimate. In this case, R determined that 9 variables at each split was mathematically sound, however, the results of tuneRF showed the model would show improvement using a value of 19 instead.

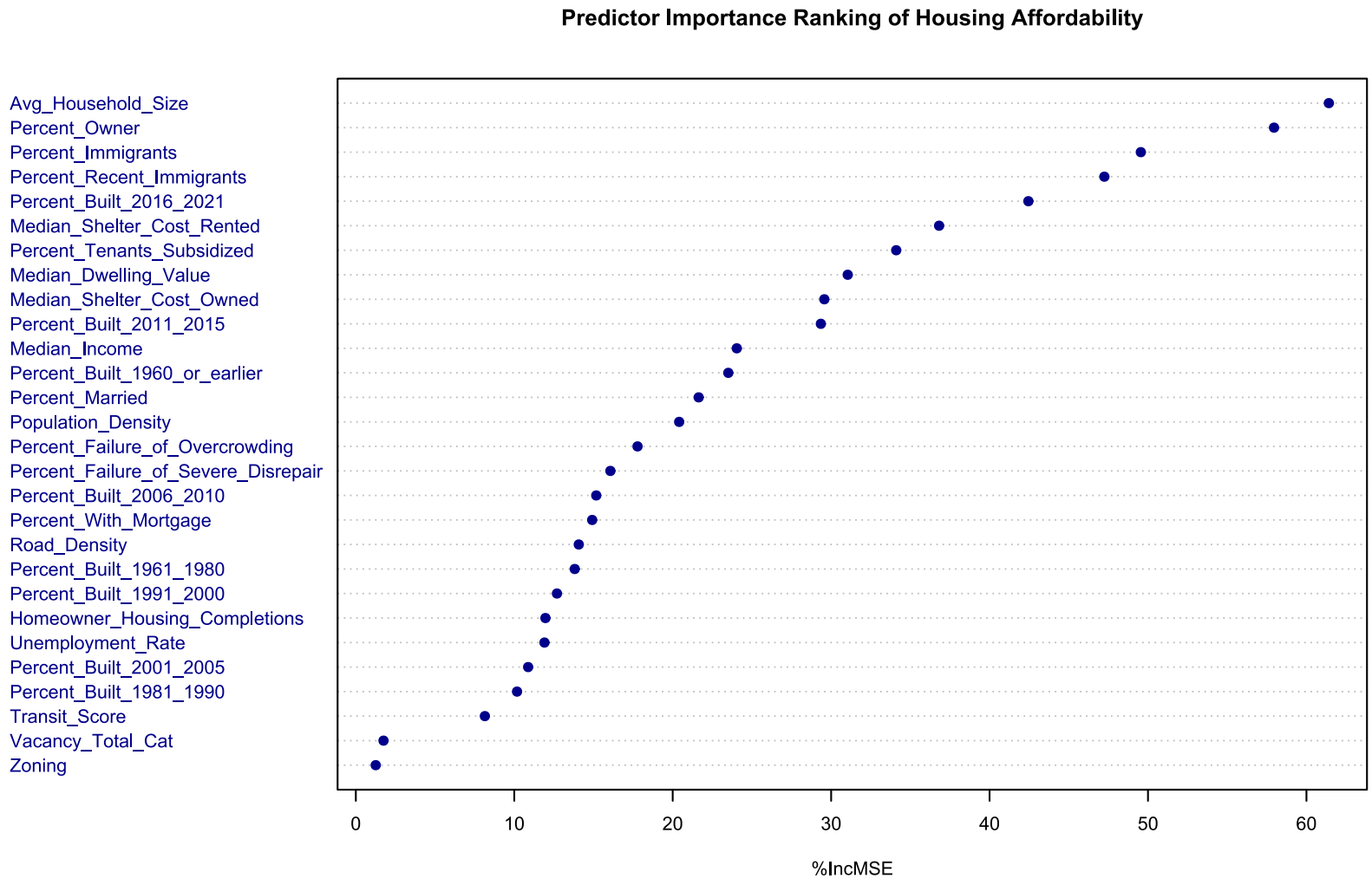


Figure 19. Variable predictor importance ranking from the random forest model using the optimal value of mtry.

```
               Type of random forest: regression
                     Number of trees: 1000
No. of variables tried at each split: 19


          Mean of squared residuals: 13.18849
                    % Var explained: 79.71
> which.min(rf_final_tuned_production$mse)
[1] 460
> sqrt(rf_final_tuned_production$mse[which.min(rf_final_tuned_production$mse)])
[1] 3.622012
```

- Variables tried at each split (mtry): 19
- Mean of squared residuals: 13.18849
- Variables explained: 79.71%
- Minimum number of trees required: 460

After tuning the model, both “Vacancy Total Cat” and “Zoning” are essentially noise to the model with their near-zero increase in mean squared error values. Therefore, these can be dropped for the final output to allow for a cleaner model to work with.

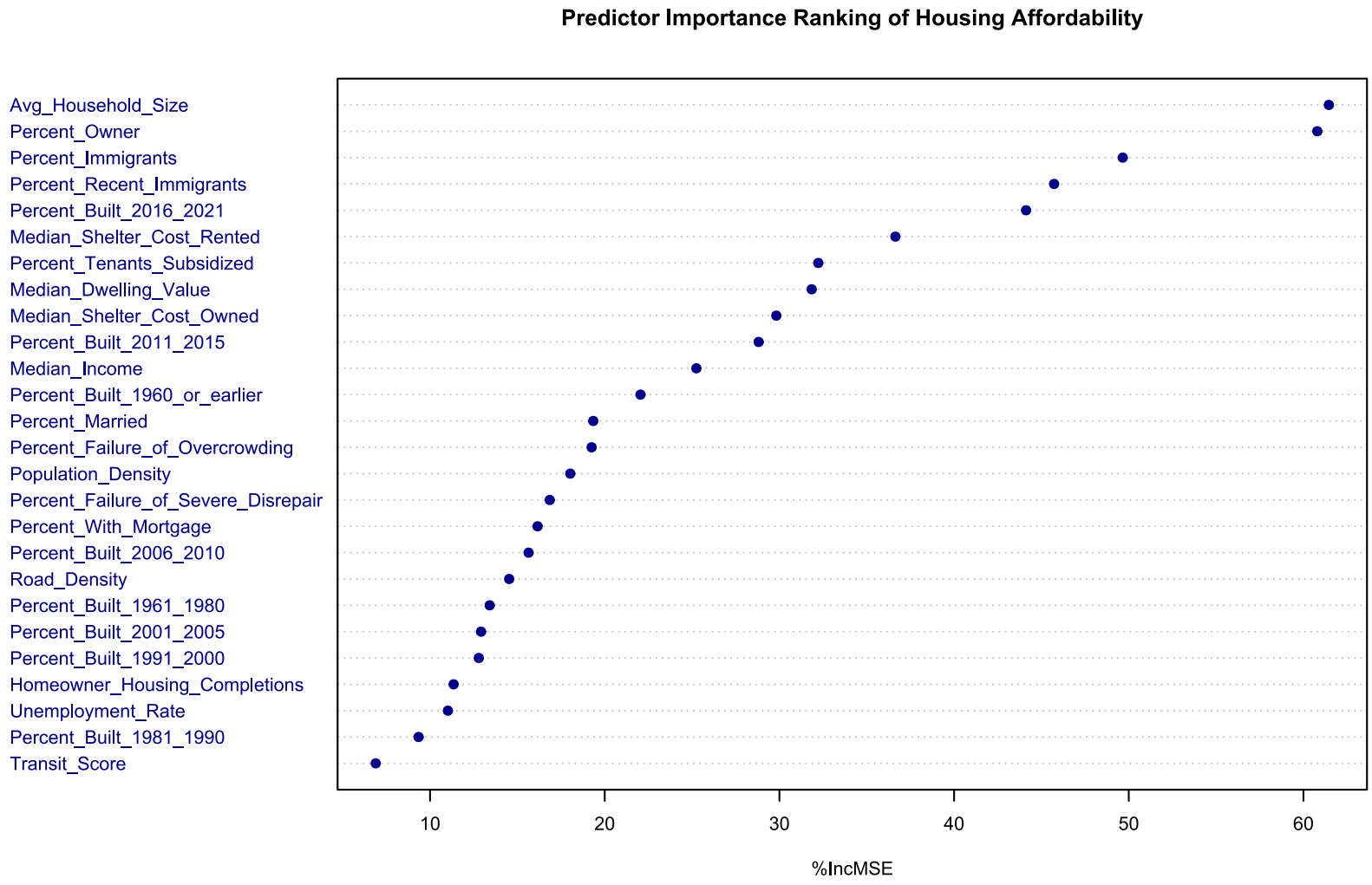


Figure 20. Variable predictor importance ranking from the random forest model with trimmed vacancy and zoning variables.

```
               Type of random forest: regression
                     Number of trees: 1000
No. of variables tried at each split: 19


          Mean of squared residuals: 13.18155
                    % Var explained: 79.72
> which.min(rf_final_tuned_production_pruned$mse)
[1] 909
> sqrt(rf_final_tuned_production_pruned$mse[which.min(rf_final_tuned_production$mse)])
[1] 3.639406
```

- Variables tried at each split (mtry): 19
- Mean of squared residuals: 13.18155
- Variables explained: 79.72%
- Minimum number of trees required: 909

With the final model complete, the next Figure shows the predictive power of each variable from this model.

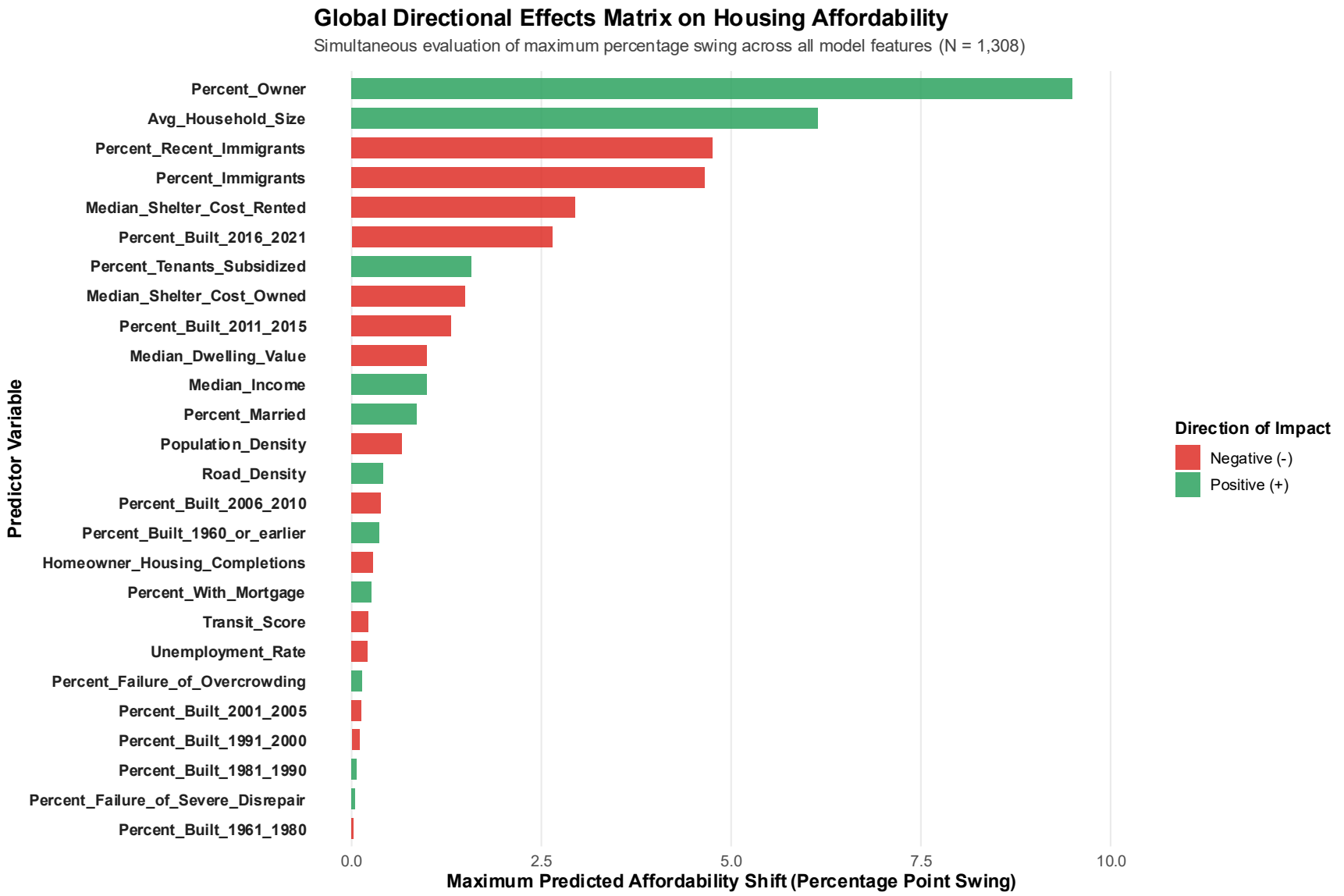


Figure 21. Global Directional Effects Matrix of Variable Predictive Power on Housing Affordability.

Figure 21 shows the final variables that were used in the random forest model, and the impacts of each. The X axis shows the percentage swing impact that each variable can affect affordability. For a positive variable like average household size, this means that at its maximum value, it can result in an almost 10% increase in housing affordability. For a negative variable like percentage of recent immigrants, at its maximum value it can result in almost 5% decrease in housing affordability. It is worth noting that the final random forest model did not end up using any categorical variables, so these values are all numerical and thus can be represented cleanly on a matrix as a percentage point swing.

3.2 Getis-Ord G_i^* Statistics

After the final random forest model was produced, G_i^* statistics were computed in R to uncover spatial patterns of affordability that went unexplained by the model.

Figure 22 represents the map of identified spatial clusters where actual affordability values did not match what was expected of the model. Figures 23 & 24 represent this data in a couple of different statistical ways, to display the distribution profiles of the complete dataset and the accuracy of the model's predictions compared to the real outcomes.

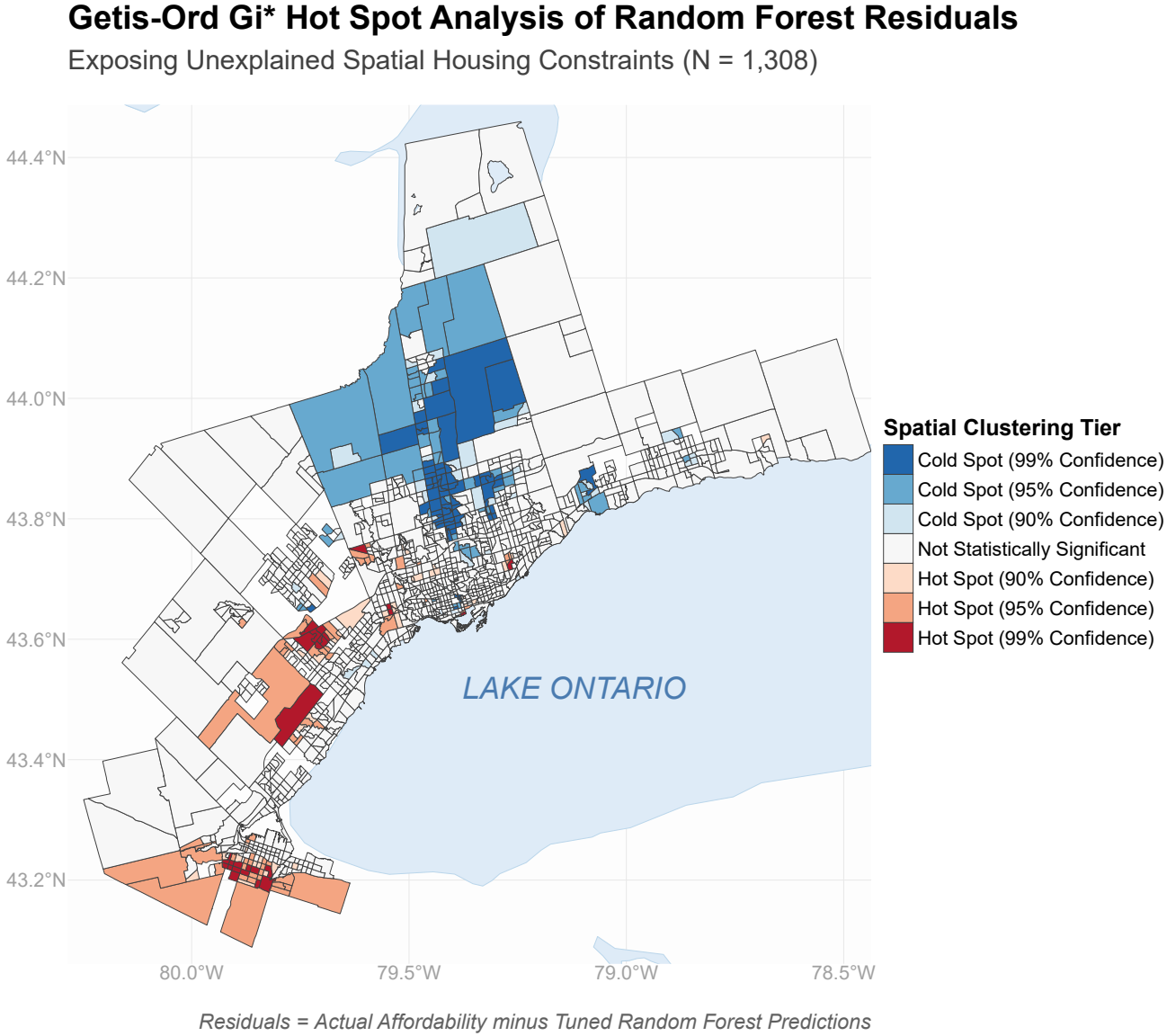


Figure 22. Getis-Ord Gi* Hot Spot Analysis of Random Forest Residuals.

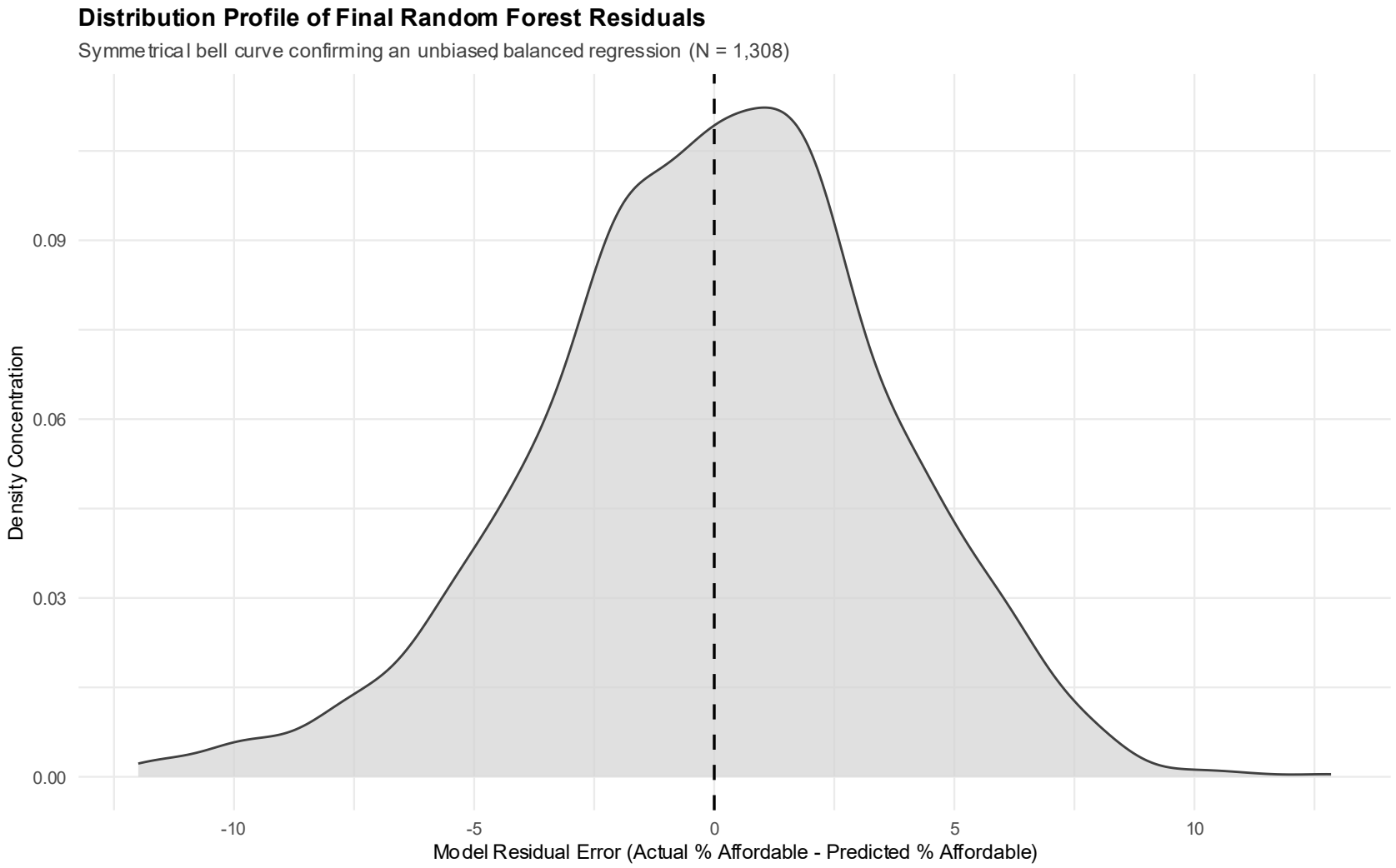


Figure 23. Distribution Profile of Random Forest Residuals.

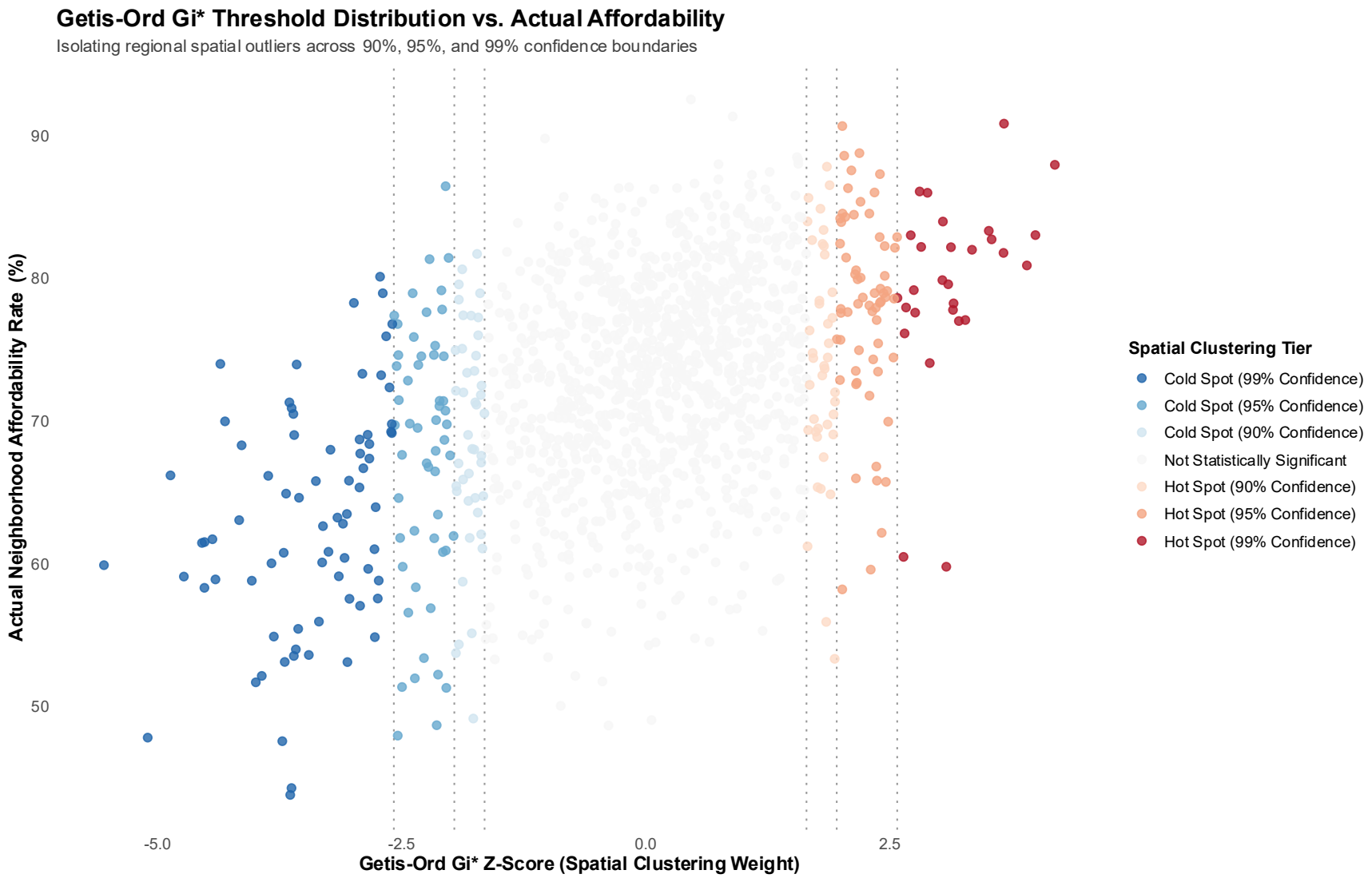


Figure 24. Getis-Ord Gi* Threshold Distribution versus Actual Affordability Plot.

The G_i^* statistics results show pockets of unexpected affordability in southern Hamilton, north Oakville/Mississauga, and a few odd pockets within Toronto, denoted by the “Hot spots” in red. Pockets of unexpected unaffordability is present in the northern parts of the Toronto CMA, primarily in York region stretching from a central core from Richmond Hill and north through Newmarket, denoted by the “Cold spots” in blue. Note that any blank spaces are census tracts that were missing data and thus were unfit for use in random forest modeling.

4. Discussion

The results of the random forest model and its variable influence matrix are generally intuitive. Perhaps the most surprising result is that while median shelter cost for both owners and renters, median income, and median dwelling value are important indicators, they are all far from the most influential. In a country like Canada where homeownership has become a retirement plan for older generations, being a homeowner tops the list of variables which are more likely to lend itself to affordable housing outcomes. The influence of homeownership is supported by the next most important variable, average household size, where larger households lead to a higher likelihood of homes being affordable in that area. People who own their own homes often have families, resulting in a larger household size, and younger people who tend to move out by themselves or with a partner/roommate are struggling more with the rising cost of housing, in both renting and homeownership models. Immigration is another key factor in assessing housing affordability for a region, with affordability trending negatively in areas with a larger share of immigrants and is slightly worse for recent immigrants, which are defined as landed immigrants between 2011-2021. As well, the discrepancy between building age brackets is alarming. Recent builds from 2016-2021 have a significant negative influence on housing affordability prediction, followed by 2011-2015 and 2006-2010. The brackets of 2001-2005, 1991-2000, 1981-1990, and 1961-1980 all have weak predictive strength, while the percentage built 1960 or earlier is the only bracket which holds somewhat significant weight in improving affordability outcomes. This result in building age, along with housing completions with an intended market of homeownership reinforces the issue of affordability not being a priority for new housing developments.

4.1 Revisiting The Three Cities Within Toronto

In John Hulchanski's longitudinal study of Toronto's income levels by census tract compared to the CMA average, there was a prediction of what 2025 would look like if the trends continued. While not a longitudinal study, it is interesting to compare what the city's current state looks like compared to the CMA average (Statistics Canada, 2023b).

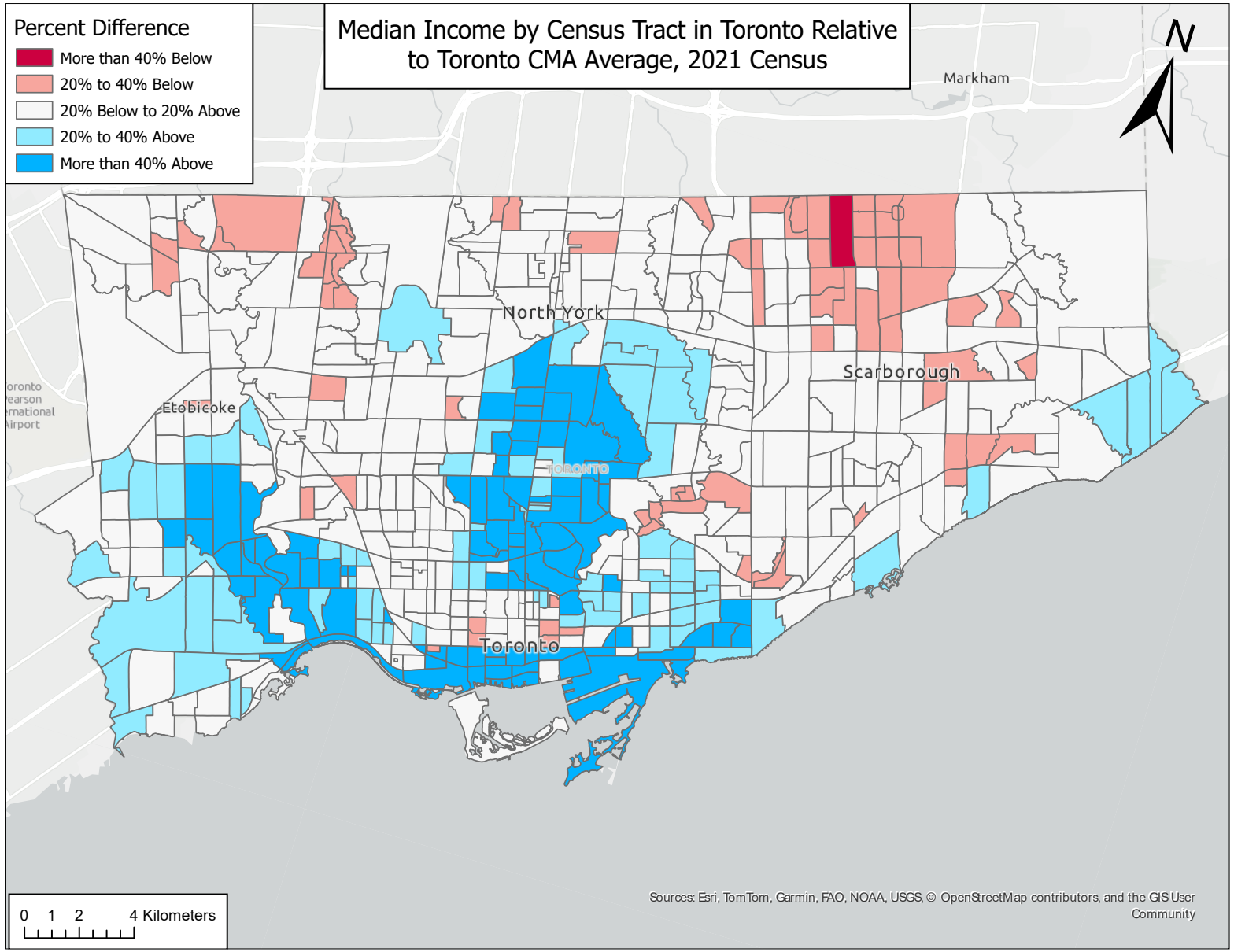


Figure 25. Median Income by Census Tract in Toronto Relative to Toronto CMA Average.

The map is interesting for a few reasons. While the general areas predicted by John Hulchanski generally hold true, particularly around the affluent areas defined as “City 1”, the city appears to be doing better than expected, at least relative to the CMA. The CMA is the key part that may be misleading, since it includes the regional municipalities of York, Peel, and parts of Durham and Halton. This map appearing average and above average for income may be due to the struggling regions outside of the city of Toronto proper, particularly the region of York as seen in the random forest modelling. This is not a longitudinal study and the city itself is not a primary concern, but it paves the way for future analysis, especially when the 2026 census of population data from Statistics Canada is available for public use.

4.2 Inclusionary Zoning

Inclusionary Zoning is a planning tool designed to incorporate affordable housing units into new developments, easing affordability pressures on the public. The concept was first introduced in 2018 by Ontario’s Liberal government, which included requirements for 5% to 10% of affordable housing units to be included per development (Javed & Jeanrie, 2026). When the current Progressive Conservative government gained power, the regulations surrounding inclusionary zoning was changed to be exclusively tied into protected major transit station areas (PTMSA), and the city of Toronto adopted their own inclusionary zoning bylaws in 2021 (Javed & Jeanrie, 2026). Under these new pieces of legislation, affordable housing requirements are tied to developments that meet the requirements of:

1. Are located within an Inclusionary Zoning Market Area;
2. Fall within an approved Protected Major Transit Station Area; and
3. Meet minimum size thresholds of 100 units and 8,000 square metres of new residential gross floor area (GFA)

(Javed & Jeanrie, 2026).

The Ontario government amended the legislation around inclusionary zoning in May 2025, now limiting affordable housing requirements to be no more than 5% of units of residential GFA, and limits the affordability terms to a maximum of 25 years (Javed & Jeanrie, 2026). On January 29, 2026, the provincial government officially paused inclusionary zoning requirements for the municipalities of Toronto, Mississauga, and Kitchener, for development applications filed on or

before July 1, 2027, through amendments to the Inclusionary Zoning Regulation under the Planning Act (Javed & Jeanrie, 2026).

Figure 26 represents a map of MTSAs within the Toronto CMA, overlaying the results from the Gi-* residuals (City of Toronto, 2025; Region of Peel, 2025; Regional Municipality of Durham, 2025; The Regional Municipality of York, 2025).

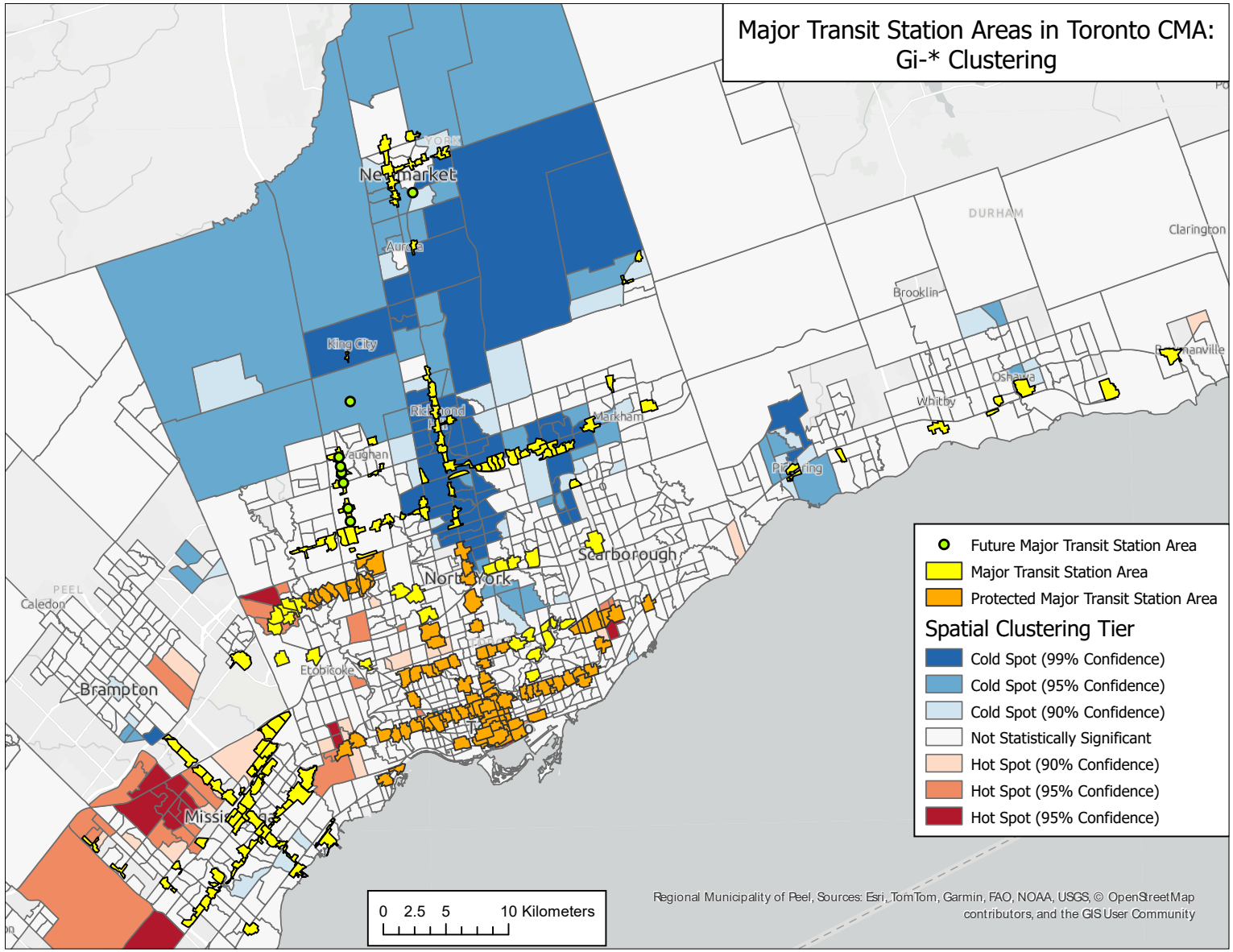


Figure 26. Major Transit Station Areas in Toronto CMA.

The MTSA map overlays some areas that have outcomes of unaffordability that the random forest model was unable to accurately predict. It can be inferred that inclusionary zoning policies would be beneficial to reducing housing stress in these areas, backed by the random forest model placing significant weight on tenants that are living in subsidized housing being a strong indicator for affordability outcomes.

4.3 Revisiting the Literature

The findings of this study broadly reinforce the theoretical perspectives presented in the literature review while providing empirical evidence regarding the spatial processes that shape housing affordability across the GTHA. Rather than identifying a single dominant determinant of affordability, the Random Forest analysis demonstrates that affordability emerges through the interaction of multiple variables. These findings support Harvey's conception of urban development as the product of interconnected spatial and economic processes.

The prominence of variables relating to housing costs, tenure, subsidized housing, household composition, and demographic change also aligns with Madden and Marcuse's argument that housing functions simultaneously as both a social necessity and a financial asset. Areas with greater concentrations of subsidized housing and more affordable housing stock generally displayed stronger affordability outcomes. This study reinforces the literature purported by scholars such as Kipfer & Sotomayor, and Lehrer et al., as neighbourhoods characterized by a higher percentage of recent builds and a higher proportion of recent immigrants tended to experience greater affordability pressures. This is a close correlation of the flawed market supply argument to solve the affordability crisis and the impacts of revitalization and Toronto's condo boom in having adverse effects on affordability. These findings suggest that housing markets continue to reproduce uneven spatial outcomes consistent with theories of commodification and housing inequality.

The results likewise support equity planning literature. Krumholz argues that planners should focus on the equitable distribution of opportunities rather than relying solely on market mechanisms. The limited predictive contribution of zoning classifications indicates that land-use regulation alone cannot adequately explain affordability patterns. Further studies are needed due to severe limitations on the effectiveness of consolidation for zoning by-laws to be used for a

random forest statistical model. It may be possible that a combination of policy factors can be used to explain affordability inequalities.

Finally, the Gi* residual analysis suggests that important place-specific processes remain beyond the explanatory power of even a strong statistical model. These unexplained clusters may reflect localized institutional decisions, redevelopment dynamics, historical investment patterns, or planning interventions such as inclusionary zoning that were not fully represented within the available datasets. Consequently, while machine-learning methods provide powerful explanatory tools, they should complement rather than replace qualitative planning analysis. Together, the findings demonstrate the value of combining critical planning theory with spatial data science to better understand and address the structural geography of housing affordability.

5. Conclusions

This study set out to answer the following research question:

How do socioeconomic, housing, accessibility, and land-use characteristics interact to explain spatial disparities in housing affordability across the Greater Toronto-Hamilton Area, and which factors most strongly predict census tract-level affordability using Random Forest spatial modelling?

As well as the following objectives:

1. To identify the socioeconomic, housing, accessibility, and land-use characteristics associated with spatial variation in housing affordability across the GTHA.
2. To develop and evaluate Random Forest regression models capable of predicting census tract-level housing affordability.
3. To assess the relative importance and direction of influence of predictor variables in explaining housing affordability outcomes.
4. To identify statistically significant spatial clusters of model residuals using Getis-Ord G_i^* statistics to highlight areas where affordability is influenced by factors beyond those represented in the model and discuss their implications for housing policy and planning.

The results demonstrate that housing affordability is best understood as the product of multiple interacting socioeconomic and housing market characteristics rather than a singular factor. The final Random Forest model explained approximately 79.72% of the variation in affordability across census tracts, indicating that neighbourhood affordability can be predicted with a relatively high degree of accuracy using publicly available spatial data. Housing tenure, household composition, immigrant status, housing costs, subsidized housing, and newer builds emerged as the strongest predictors, while older housing builds and housing suitability categories contributed comparatively little additional explanatory power. Numerous variables were dropped throughout the evolution of each random forest model due to producing additional noise and having seemingly zero or unpredictable correlations to housing affordability outcomes. These included zoning characteristics and vacancy rates.

These findings satisfy the first objective by identifying the principal variables associated with housing affordability across the GTHA. The second objective was achieved through the successful development and evaluation of Random Forest regression models that substantially

outperformed individual decision trees. The third objective was addressed through variable importance analysis, which revealed both the relative predictive strength and direction of influence of individual variables. Finally, the application of Getis-Ord G_i^* statistics achieved the fourth objective by identifying statistically significant spatial clusters where affordability outcomes remained insufficiently explained by the model, suggesting the presence of additional institutional, historical, or policy influences.

More broadly, the findings in this study reinforce critical urban planning literature that views housing affordability as a structural outcome of economic, political, and spatial processes rather than simply a question of housing supply. While this is partly due to the difficulty in consolidating zoning data in a way that is cohesive for a model to analyze, the limited explanatory contribution of zoning as a variable suggests that planning interventions should be considered alongside broader social, fiscal, and housing policies. Machine-learning approaches such as Random Forest provide planners with valuable decision-support tools capable of identifying complex nonlinear relationships while remaining interpretable enough to inform evidence-based policy.

Future research should incorporate newer Census data, longitudinal analyses, and additional policy variables. This includes inclusionary zoning, affordable housing funding programs, purpose-built rentals, transit-oriented development policies, and rent control policies. A longitudinal study would allow for a greater understanding of housing affordability trends over time, much like what was outlined in Hulchanski's Three Cities.

While the reason behind the existence of the affordability/unaffordability pockets identified through Gi* analysis is beyond the scope of this study, the spatial patterns uncovered can lend itself to highly interesting future studies. Despite a near 80% variance explained value produced by the random forest model, these areas are distinct and clear, representing outcomes that went unexplained by census data. A deeper dive into the housing policy, historical, environmental and institutional factors of these regions could explain why this phenomenon occurred. It is possible that one or more of the many possible policy variables that was not part of this study could be the explanation for why these neighbourhood regions did not fit with the rest of the dataset.

Together, these findings demonstrate that integrating GIS, explainable machine learning, and critical planning theory offers a valuable framework for advancing housing affordability research and informing more equitable planning practice.

References

August, M., & Walks, A. (2018). Gentrification, suburban decline, and the financialization of multi-family rental housing: The case of Toronto. *Geoforum*, *89*, 124–136. https://doi.org/10.1016/j.geoforum.2017.04.011

Bivand, Roger S., Edzer Pebesma, and Virgilio Gómez-Rubio. 2013. *Applied Spatial Data Analysis with R, Second Edition*. Springer, NY. https://asdar-book.org/.

Bivand, Roger, and David W. S. Wong. 2018. "Comparing Implementations of Global and Local Indicators of Spatial Association." *TEST* 27 (3): 716–48. https://doi.org/10.1007/s11749-018-0599-x.

Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (Eds.). (1984). *Classification and regression trees*. Wadsworth & Brooks/Cole.

Caledon Zoning By-law. (2025). [Dataset].

City of Brampton. (2025). *Brampton Bus Stops* [Dataset].

City of Brampton. (2026). *Brampton Official Plan Zoning* [Dataset].

City of Burlington. (2023). *Burlington Zoning ByLaw* [Dataset].

City of Burlington. (2026). *Burlington Transit Bus Stops* [Dataset].

City of Hamilton. (2026a). *Hamilton Zoning By-law Boundary* [Dataset].

City of Hamilton. (2026b). *HSR Bus Stops* [Dataset].

City of Markham. (2024). *Markham Zoning* [Dataset].

City of Mississauga. (2021). *City of Mississauga Zoning Bylaw* [Dataset].

City of Mississauga. (2026). *MiWay Bus Stops* [Dataset].

City of Oshawa. (2017). *Oshawa Zoning* [Dataset].

City of Pickering. (2024). *Pickering Zoning* [Dataset].

City of Richmond Hill. (2026). *Richmond Hill CZBL Schedules* [Dataset].

City of Toronto. (2024). *City of Toronto Zoning By-law 569-2013* [Dataset].

City of Toronto. (2025). *Toronto MTSA* [Dataset].

City of Vaughan. (2026). *Vaughan Zoning1_21* [Dataset].

CMHC. (2026). *Starts and Completions Survey and Market Absorption Survey Methodology*.

Friedl, M. A., & Brodley, C. E. (1997). Decision tree classification of land cover from remotely sensed data. *Remote Sensing of Environment*, *61*(3), 399–409.

Getis, A., & Ord, J. K. (1992). The Analysis of Spatial Association by Use of Distance Statistics. *Geographical Analysis*, *24*(3), 189–206. https://doi.org/10.1111/j.1538-4632.1992.tb00261.x

Government of Ontario. (2026a). *GO Train Stations* [Dataset].

Government of Ontario. (2026b). *Ontario Road Network* [Dataset].

Greenwell, Brandon M. 2017. "pdp: An R Package for Constructing Partial Dependence Plots." *The R Journal* 9 (1): 421–36. https://doi.org/10.32614/RJ-2017-016.

Hackworth, J., & Moriah, A. (2006). Neoliberalism, Contingency and Urban Policy: The Case of Social Housing in Ontario. *International Journal of Urban and Regional Research*, *30*(3), 510–527. https://doi.org/10.1111/j.1468-2427.2006.00675.x

Harvey, D. (Ed.). (2009). *Social justice and the city* (Rev. ed). University of Georgia Press.

Harvey, D. (2012). The urban roots of financial crises: Reclaiming the city for anti-capitalist struggle. *Socialist Register*, *48*, 1–35.

Hulchanski, J. D. & University of Toronto. (2010). *THE THREE CITIES WITHIN TORONTO - Income Polarization Among Toronto's Neighbourhoods, 1970-2005*. University of Toronto Centre for Urban and Community Studies.

Javed, S., & Jeanrie, A. (2026, February 5). Inclusionary Zoning in Ontario: Ontario Moves to Pause Inclusionary Zoning Obligations. *Bennett Jones*.

https://www.bennettjones.com/Insights/Blogs/Ontario-Moves-to-Pause-Inclusionary-Zoning-Obligations

Kipfer, S., & Sotomayor, L. (2024). Housing beyond land rent?: A critique of market housing solutionism. *Radical Housing Journal*, *6*(1), 33–61. https://doi.org/10.54825/GQAU9087

Krumholz, N. (1986). CITY PLANNING FOR GREATER EQUITY. *Journal of Architectural and Planning Research*, *3*(4), 327–337.

Krumholz, N., & Hexter, K. W. (Eds.). (2019). *Advancing Equity Planning Now*. Cornell University Press. https://doi.org/10.7591/9781501730399

Lehrer, U., Keil, R., & Kipfer, S. (2010). Reurbanization in Toronto: Condominium boom and social housing revitalization. *disP - The Planning Review*, *46*(180), 81–90. https://doi.org/10.1080/02513625.2010.10557065

Leon, S., & Iveniuk, J. (2021). *Widening inequities: Long-term housing affordability in the Toronto Census Metropolitan Area 1991-2016*. Wellesley Institute.

Liaw, Andy, and Matthew Wiener. 2002. "Classification and Regression by randomForest." *R News* 2 (3): 18–22. https://CRAN.R-project.org/doc/Rnews/.

Madden, D., & Marcuse, P. (2016). *In Defense of Housing: The Politics of Crisis*. Verso.

Marcuse, P. (1978). Housing Policy and the Myth of the Benevolent State. *Social Policy*, *8*(4), 21–26.

Massicotte, Philippe, and Andy South. 2026. *rnaturalearth: World Map Data from Natural Earth*. https://doi.org/10.32614/CRAN.package.rnaturalearth.

Metrolinx. (2026). *Go Transit GTFS* [Dataset].

Milborrow, Stephen. 2026. *rpart.plot: Plot "rpart" Models: An Enhanced Version of "plot.rpart"*. https://doi.org/10.32614/CRAN.package.rpart.plot.

Municipality of Clarington. (2026a). *Clarington Zoning By-law 84-63* [Dataset].

Municipality of Clarington. (2026b). *Clarington Zoning By-law 2005-109* [Dataset].

Oakville Zoning By Law 2009-189. (2024). [Dataset].

Oakville Zoning By Law 2014-014. (2024). [Dataset].

Pebesma, Edzer. 2018. “Simple Features for R: Standardized Support for Spatial Vector Data.” *The R Journal* 10 (1): 439–46. https://doi.org/10.32614/RJ-2018-009.

Pebesma, Edzer, and Roger Bivand. 2023. *Spatial Data Science: With applications in R*. Chapman and Hall/CRC. https://doi.org/10.1201/9780429459016.

Pebesma, Edzer, and Roger S. Bivand. 2023. *Spatial Data Science with Applications in R*. Chapman & Hall. https://r-spatial.org/book/.

Prasad, A. M., Iverson, L. R., & Liaw, A. (2006). Newer classification and regression tree techniques: Bagging and random forests for ecological prediction. *Ecosystems*, *9*(2), 181–199.

Region of Peel. (2025). *Peel Region MTSA* [Dataset].

Regional Municipality of Durham. (2020). *Transit Bus Boarding Points* [Dataset].

Regional Municipality of Durham. (2025). *Durham Region MTSA* [Dataset].

Roger Bivand. 2022. “R Packages for Analyzing Spatial Data: A Comparative Case Study with Areal Data.” *Geographical Analysis* 54 (3): 488–518. https://doi.org/10.1111/gean.12319.

Statistics Canada. (2023a). *2021 Census of Population* [(Table)].

Statistics Canada. (2023b). *Census Profile* (Nos. 98-316-X2021001) [(Table)].

The Regional Municipality of York. (2024). *Bus Stops from GTFS* [Dataset].

The Regional Municipality of York. (2025). *York Region MTSA* [Dataset].

Therneau, Terry, and Beth Atkinson. 2026. *rpart: Recursive Partitioning and Regression Trees*. https://doi.org/10.32614/CRAN.package.rpart.

Toronto Transit Commission. (2026). *TTC Routes and Schedules* [Dataset].

Town of Ajax. (2021). *Ajax Zoning* [Dataset].

Town of Aurora. (2025). *Aurora Lots Zoning* [Dataset].

Town of East Gwillimbury. (2025). *East Gwillimbury Zoning* [Dataset].

Town of Georgina. (2021). *Town of Georgina Zoning* [Dataset].

Town of Halton Hills. (2025). *Halton Hills Zoning By-law 2010-0050* [Dataset].

Town of Milton. (2022). *Milton Rural Zoning By-law 144-2003* [Dataset].

Town of Milton. (2024). *Milton GTFS* [Dataset].

Town of Milton. (2026). *Milton Urban Zoning By-law 016-2014* [Dataset].

Town of Uxbridge. (2023). *Uxbridge Zoning By-law 81-19* [Dataset].

Town of Whitby. (2024). *Town of Whitby Zoning* [Dataset].

Town of Whitchurch-Stouffville. (2026). *Stouffville Planning Zoning* [Dataset].

Township of King. (2025). *Township of King Zoning and Conservation Layers* [Dataset].

University of Toronto. (2022). *Canadian Census Analyser (CHASS)* [Dataset].

Whitzman, C., & Slater, T. (2006). Village Ghetto Land: Myth, Social Conditions, and Housing Policy in Parkdale, Toronto, 1879–2000. *Urban Affairs Review*, *41*(5), 673–696. https://doi.org/10.1177/1078087405284673

Wickham, Hadley, Mara Averick, Jennifer Bryan, et al. 2019. “Welcome to the tidyverse.” *Journal of Open Source Software* 4 (43): 1686. https://doi.org/10.21105/joss.01686.

Wickham, Hadley, Lionel Henry, Thomas Lin Pedersen, T Jake Luciani, Matthieu Decorde, and Vaudor Lise. 2025. *svglite: An “SVG” Graphics Device*. https://doi.org/10.32614/CRAN.package.svglite.

Zapata, M. A., & Bates, L. K. (2015). Equity Planning Revisited. *Journal of Planning Education and Research*, *35*(3), 245–248. https://doi.org/10.1177/0739456X15589967