CONCORDIA UNIVERSITY

DEPARTMENT OF ECONOMICS

Air Pollution and the Canadian Real Estate Market: Addressing Regressor Endogeneity using Meteorological Data

M.A. RESEARCH PAPER

Student's Name: Nigel McKernan
Student ID: 40052356
Date Submitted:

Supervisor: Dr. Xintong Han

Department of Economics
Signature page

This is to certify that the M.A. Research Paper prepared

By: Nigel McKernan ID#: 40052356

Entitled: Air Pollution and the Canadian Real Estate Market: Addressing Endogeneity using Meteorological Data

and submitted in partial fulfillment of the requirements for the degree of

Master of Arts (Economics)

complies with the regulations of the University and meets the accepted standards with respect to originality and quality.

Signed by the final examining committee:

Supervisor: 韩信甫 Xinfu Han.
Date: 08/07/2020

Examiner:
Date: 8/6/20

Approved by Graduate Program Director: _____
Date: _____

Grade Sheet prepared by Graduate Program Assistant (given to Supervisor):
Date: _____

Air Pollution and the Canadian Real Estate Market: Addressing Endogeneity using Meteorological Data

Nigel McKernan

Abstract

The hedonic pricing model is commonly used in addressing the issue of air pollution's effect on the real estate market, as it allows for single-variable isolation in inferring a variable's effect on house prices. However, pollution as a variable is found to be *endogenous* per the literature; researchers often omit various factors that can influence local particulate matter concentrations that have little impact on house prices themselves. This usually leads to researchers deriving an undesirable sign (positive) for their 'pollution' regressor. The objective of this analysis is to incorporate two meteorological parameters (wind direction and wind speed) as instrumental variables (IV) to address the endogeneity of air pollution, while keeping other socioeconomic data constant. The dataset used is for 8 Census Metropolitan Areas (CMA) in Canada, across an unbalanced panel of approximately 16 years (depending on the city), with observations following a monthly frequency. Additionally, as the various types of meteorological data did not conform with the desired monthly data-recording frequency, temporal disaggregation is implemented to express all data in a monthly frequency. Implementing a panel-data instrumental variable approach results in a *negative* coefficient of significant magnitude for the pollution regressor. The resulting model demonstrates a highly statistically significant inverse correlation between real estate prices and a surrogate measure (concentrations of particulate matter of 2.5 microns average diameter) of 'local air pollution'.

Table of Contents

List of Tables

List of Figures

I. Introduction

As science drives the world's understanding of air pollution forward, econometric (and many other types of) analysts look to answer the increasingly imperative question of exactly *how much* the public values the impact of pollution. Researchers favour the use of a hedonic pricing model vis-a-vis real estate prices. This model is constructed by the interaction between housing prices and pollutant concentrations of one particular area or city, assuming all observable dwelling, neighbourhood, or city-level characteristics are controlled for. However, a prevalent issue that is not commonly addressed in the literature is the endogeneity of pollution. Hong (2015) was among the first to implement a hedonic model incorporating pollution in the context of the Canadian metropolitan real estate market. However, Hong derived undesirable (positive) signs for their pollution regressor, despite multiple attempts to improve their model specification and construction. Hong eventually derived a negative sign for their pollution regressor after incorporating lagged values of the dependent variable as an exogenous regressor. Moreover, no attempt to address any kind of regressor endogeneity was made in their analysis.

Additionally, Chay and Greenstone (2005) find pollution to be an *endogenous* regressor; there are often factors which affect local air pollution concentrations that are *not* correlated with house prices themselves. As such, a major motivation for this paper is to improve upon Hong's analysis by addressing the potential endogeneity in the pollution regressor. This is achieved by implementing a panel-data, instrumental-variable (IV) approach, using a two-stage-least-squares (2SLS) regression. Chay and Greenstone's instrument was a "non-attainment status" label assigned by the United States Federal

Government to counties which polluted over a specified air-quality threshold(s) within the observed year. In Canada, however, no equivalent standards are in effect for the provinces, let alone metropolitan areas, and as such, this analysis requires different instrumental variables.

Bondy et al. (2018) also find pollution to be an endogenous regressor in their analysis of establishing the connection between crime rates and air pollution in inner-city boroughs in London. They achieved this by using meteorological data as instrumental variables to supplant their pollution regressor; local air-pollution concentrations were correlated with local meteorological factors such as wind speed, wind direction, relative humidity, and precipitation which could then be used as instruments for pollution.

Thus, this paper's approach is to combine the methods laid out by Hong, Bondy et al. and Chay and Greenstone to address the endogeneity of air pollution. Specifically, this paper combines Hong's analysis of Canada's metropolitan real estate market with Bondy et al. (2018)'s use of meteorological data as instrumental variables, using wind direction and wind speed in a panel-data setting. Additionally, this paper incorporates more sociological variables as exogenous regressors compared with Hong's analysis by temporally disaggregating data derived from Statistics Canada from an annual frequency to a monthly frequency.

When implementing a 2SLS model, the use of wind direction and wind speed as instrumental variables results in a *negative* sign for the pollution regressor with significant magnitude, without having to explore including lagged values of certain variables as additional regressors via a Dynamic GMM approach, à la Arellano & Bond (1991).

II. Literature Review

i. Hedonic Models and Pollution

The economic interpretation of the hedonic pricing model is that the various characteristics of a particular household, and characteristics pertaining to its respective neighbourhood, or even city, can be partially differentiated to infer a marginal effect on the value of a dwelling, if any of its characteristics or that of its region change by one unit. For example, the marginal effect of adding an additional bedroom to a dwelling should affect its value in some direction and magnitude, keeping constant all other observed dwelling or neighbourhood-specific characteristics. Rosen (1974) was the first to put this into a specific economic meaning, per Chay and Greenstone (2005). However, several identification problems were pointed out following the publications of several papers, particularly Rosen's, tackling hedonic pricing models when incorporating pollution into their regressions. Chief among them were omitted variables; Small (1975) criticized Rosen's analysis for having omitted unobserved neighbourhood characteristics that would be correlated with pollution, thus leading to specification issues when attempting to capture the true impact of pollution on housing prices. Another important issue is self-selection bias: individuals or households that have lower disutility to pollution might self-select into areas with higher pollution levels to capitalize on lower house prices. This introduces bias in the effects of pollution on house prices and makes estimating the marginal willingness-to-pay (MWTP) curve more difficult, as taste cannot be assumed to be homogeneous across all individuals or households.

Chay and Greenstone's robust analysis of the connection between house prices and pollution used a novel instrumental variable approach, along with a variable coefficients model to combat endogeneity and self-selection biases, respectively. They collected a rich dataset containing data on numerous county-level air-quality variables, including the total suspended particulate (TSP) parameter for several counties in the United States, via a Freedom of Information Act request, for the years 1970 and 1980. In trying to combat the main issues pointed out by Small, Chay and Greenstone utilize an IV approach to circumvent the endogeneity bias: counties that polluted above a certain amount mandated by the United States government as being tolerable were designated as having a "non-attainment" status. They argue that this designation has a direct effect on air pollution, which is not correlated with house prices or taste. When they included county-level covariates and flexible forms of these covariates (i.e. quadratic and cubic forms), they found a negative association between house prices and air pollution. Chay and Greenstone also incorporate different approaches to explore the extent to which tastes among the different counties were heterogeneous in their preference to pollution. They accomplished this by complementing their instrumental difference-in-differences approach with a variable coefficients model, in which each county has its own coefficient on the marginal effect of pollution on house prices. By implementing a difference-in-differences approach, and utilizing instrumental variables, Chay and Greenstone handily addressed the issue of omitted variable bias by eliminating time-invariant unobserved factors by first-differencing, and accounting for variable endogeneity by using "non-attainment status" as an instrument for pollution (measured in $\mu g/m^3$ of $PM_{2.5}$). The other main issue pointed out by Small is the heterogeneity in preference (taste) with regard to pollution and house prices;

those who have a lower disutility to pollution will move to areas with higher pollution to capitalize on lower house prices. To account for this, they implement a variable-coefficients model, whereby each county obtains its own coefficient on the marginal effect of an increase in total suspended particulate matter on house prices. They found that the average non-random sample of a county's population reflects a negative correlation with regard to pollution and house prices when accounting for regional fixed effects and covariates taken from the census (including quadratic transformations) for the years Chay and Greenstone were able to obtain data. In other words, the coefficient for pollution was relatively poolable. The range to which this coefficient varied across counties was not found to be large, and was shown to be negative.

In his analysis of pollution affecting Canadian house prices, Hong (2015) acknowledges the same issues echoed by Chay and Greenstone and utilizes a longitudinal model by collecting data describing annual resale prices of homes, averaged over the level of the "Census Metropolitan Area" (hereafter CMA). Statistics Canada defines a Census Metropolitan Area as an "area consisting of one of more neighbouring municipalities situated around a core. A CMA must have a total population of at least 100,000 of which 50,000 or more live in the core." Hong utilizes two indices to capture annual average house prices: the New Housing Price Index (NHPI) from CANSIM, and the Teranet-National Bank House Price Index (THPI). The THPI index gave more robust results (per Hong) because this index also captures resale prices of homes, and not just initial sale price. However, due to utilizing annual data, they did not have many explanatory covariates to control for at a fine geographic level, utilizing only the unemployment rate, average total income, population at time $T = t$, and through further model specification, pollution at

time $T = t - 1$, as these are all available on an annual basis from Statistics Canada. Hong obtains pollution data from Environment Canada, where total suspended particle readings are collected by 64 federal monitoring stations across Canada. Through various iterations of his model, such as fixed-effects, random effects, and first-differencing, Hong finds a negative association between pollution and house prices at the geographical scale of the CMA, although only through reliance on a lagged-time variable, as they acknowledge that the effects of pollution likely aren't to be felt immediately.

Expanding the context of Canadian hedonic models, Li et al. (2006) conduct a similar study to Hong's but omit any focus on pollution or another type of environmental externality that might influence the average/median house prices. And while they do compare and contrast different functional forms of their model, including a semi-log, log-log, and a linear model with a Box-Cox transformation applied to the dependent variable, they model only one cross-section in the Ottawa-Gatineau region. As such, they are unable to control for various unobserved characteristics of the prices of these houses in general, although they do collect more data at the neighbourhood/borough scale than this paper, because they were able to obtain exclusive data generated by the Multiple Listing Service (MLS). They find that conducting a Box-Cox transformation rejects the functional form of a linear model, and variants thereof, including logs in the model. The Box-Cox transformation, while correcting for non-normality in the transformed variable, changes the interpretation of this variable in terms of its units. In addition, the dependent variable need not have a normal distribution, and needs only that the errors from our model conform closely to a normal distribution.

Chan (2014) focuses on the West Vancouver real estate market, focusing on incorporating spatial data within the city of West Vancouver to differently weigh the observations. They find that Ordinary-Least-Squares (OLS) overestimates real estate values by not being able to incorporate spatial effects into considerations of real estate values, leading to a geographically-weighted regression model, resulting in the best model for out-of-sample prediction of real estate prices.

ii. Wind Direction & Speed as Instruments

In the previous section, it was noted that endogeneity of pollution is a persistent issue in deriving accurate measurements of the effects of local pollution on house prices. Chay and Greenstone utilized the 1974 "non-attainment" designation of over-polluting counties in the United States between the years 1970 and 1980 to instrument for pollution. However, the Canadian government has not set equivalent standards for air pollution at the CMA level, and as such, there is no equivalent instrument, therefore an alternate variable is needed to combat the issue of endogeneity.

Bondy et al. (2018) derive a link between air pollution and crime rates in inner-city boroughs of London, England, by utilizing wind direction as an instrument for air pollution. In addition to exploiting a panel-data structure, and comparing fixed and random-effects models to eliminate time-invariant unobserved factors, they also use wind direction to combat the time-varying factors as well, finding a relevant and robust instrument to establish a contemporaneous link between crime rates and air pollution in London. They cite other sources conducting similar research in Chicago and Los Angeles, noting the work of Anderson (2015) and Deryugina et al. (2016) that find similar positive links between crime rates and air pollution, and with the use of daily wind direction as an appropriate

instrument for measured air-pollution parameters. However, in Bondy et al. (2018), London has the infrastructure, and the environmental monitoring capability, to collect TSP readings across several boroughs of London, with wind direction found to be heterogeneous across London. This localized monitoring capacity allowed borough-level data to be utilized when constructing their model. In the case of this paper, as multiple metropolitan areas will be compared, the availability of comparing this heterogeneity of wind and environment data is not feasible.

iii. Temporal Disaggregation

It is a frequent issue in economics that the recording frequency of the temporal data available does not meet the frequency needed to make robust inferences based on large-sample asymptotic properties. High-frequency data also allow the researcher to account for the variations in the dataset that occur at that lower monitoring-frequency level, whereas those variations often disappear in being aggregated in a higher (i.e. longer) frequency level. For example, fluctuations of TSP readings and other environmental variables vary minute to minute, hour to hour, day to day, let alone, month to month. This variation in the data is often completely absorbed when making the aggregation to a longer frequency that conforms with the recording frequency of other potentially related data. Garrett (2002) compares the differing ways in which data available at different monitoring frequencies, due to being aggregated, or disaggregated, can affect the inferential capabilities of regression estimators, and their efficiency. They summarise that the Residual Sum of Squares (RSS) of regressions conducted on aggregated data often "can be larger or smaller than the sum of RSS of less aggregated regressions". This is found mainly in the correlation between the residuals of the regressions of the disaggregated data. If these residuals are

found to be highly correlated with other regressions at the same level of data disaggregation, then standard errors are affected. Consequently, the overall R-squared is deflated at the aggregated-data level. Often the costs associated with high-quality data collection at an elevated frequency become prohibitive for many agencies to conduct. As such, the data are usually collected at the most economically feasible rate, depending on the party's balancing of technical needs and pragmatism.

For an example relating to this analysis, Statistics Canada collects a substantial amount of information in Census Metropolitan Areas (hereafter CMA) across Canada at many different time frequencies, including quarterly, annually, monthly, semi-annually, and so on. However, as addressed in the following Data Collection section, the environmental data collected for this project (i.e. wind direction and speed, and respirable particulate matter with a mean diameter of 2.5 microns) are available at a very fine temporal frequency from Environment Canada. Air-quality data from the nation-wide monitoring network are available at an hourly frequency, (with some missing observations), and wind direction and speed data, as well as other environmental data (like precipitation) are also available at an hourly frequency, if not, daily. These time frequencies contrast with the covariates used in this model that pertain to the demographic and socioeconomic factors of census metropolitan areas. These latter variables are available usually only at an annual frequency, although sometimes monthly data can be found. This creates a challenge in defining how compromise will best be established between: (i) how aggregation to a lower-frequency time-averaging period will absorb (and obscure) the fine variations in the high-frequency datasets, on the one hand, and (ii) how best to disaggregate lower-frequency data (e.g. socioeconomic factors, like median family income) to a finer

frequency that will perform well in statistical inference, on the other. The compromise established in this paper will be disaggregate or aggregate data to a monthly frequency, which corresponds with the data-recording frequency used in the Teranet House Price Index and CANSIM's National House Price Index datasets. The environmental data (e.g. $PM_{2.5}$ concentrations, wind direction, and wind speed) were thus aggregated to a monthly time step by averaging the observations recorded over the days/hours for the period of record, omitting any missing observations. All demographic and socioeconomic data were disaggregated from an annual level to the same monthly frequency. This was accomplished using the Denton-Cholette temporal disaggregation algorithm, based on methods proposed by Denton (1971) and Dagum and Cholette (2006). This algorithm, along with others such as Jacobs (1994), and Wei and Stram (1990), perform disaggregation using purely mathematical methods, having no regard to the variations from the economic or sociological factors contributing to variability that might occur at that higher frequency, as is usually addressed in methods used by Chow-Lin (1971), and Litterman (1983). Both types of algorithms rely on an "indicator series", a time series which serves as a "guide"; this is a high-frequency time series that contains variations in its data that the lower-frequency dataset can emulate. This is usually done by estimating an Autoregressive of Order 1 (AR(1)) coefficient to impute the observations of the higher-frequency time series. This can present an issue, where a dataset recording at the desired higher frequency, that is also correlated with the lower frequency data, might not be available to perform the imputations. This is where the methods proposed by Denton (1971) can perform these imputations, as they do not rely on available and sufficiently correlated empirical data series, rather they estimate an autoregressive coefficient, where the indicator series is only

a constant; usually a vector of ones. This allows the higher-frequency observations to be imputed, but at a rate that is relatively smooth, and does not deviate much from the overall structure and movements from the original higher-level time series data. Of course, if an empirical indicator series is available, and if it is sufficiently correlated with the data needed to be disaggregated, that will emulate the empirical movements in the data that likely did occur but, because the data collection did not occur at this frequency, could not be observed. In regard to this project, of the data collected by Statistics Canada to be used as control covariates, most are available at only an annual frequency. Further, and unfortunately, data available at a monthly frequency concerning the geography of Census Metropolitan Areas would not be appropriate to perform these imputations, or were not available for enough of the Census Metropolitan Areas. In addition, when selecting indicator series with which to perform imputations, with complex time-series data like median household income, there is little consensus as to what series can serve as a good basis to perform the imputations. As well, it may be one of many, and traditional goodness-of-fit measures might not be representative of what the best imputations or forecasts are. Muller-Kademann (2014) derives a methodology to validate the strength of temporal disaggregation imputations by testing the internal consistency of the model itself using what Muller-Kademann calls a "cloud chamber" approach, similar to how physicists measure the fallout of undetectable particles after exposure to radioactive decay. This is quite novel, as Muller-Kademann notes, as conventional methods to validate the temporal disaggregation rely on external validations. Further, in comparing imputation based on an empirical dataset correlated with the time series needing to be imputed, with imputation based on a constant, even when using a strong indicator series, the two time series might

drift apart from each other, or follow movements that do not conform with mutual consistency. As such, imputation in this project will rely on imputation by estimating the AR(1) coefficient based on a constant-vector for the relevant time series.

III. Data Sources and Collection

As was previously noted, the three main categories of data constituting this analysis consist of the real-estate data for residential property values derived from the Teranet House Price Index (THPI) and the National House Price Index by CANSIM (NHPI), Environment Canada's historical environmental data that will serve as the main (endogenous) independent variable and the instruments that will be used to account for endogeneity, and other controlling covariates (constituting demographic and socioeconomic factors) from Statistics Canada. Each of these three types of data, as described previously, is available in differing frequency of collection from one to each other. The THPI and NHPI data are available on a monthly basis, data from Environment Canada are available on a daily or hourly basis, and the covariate data from Statistics Canada are mostly available on an annual basis. As was detailed at the end of the previous section, the latter two categories of data will be aggregated and disaggregated, respectively, to a monthly frequency to match the frequency of the THPI/NHPI housing indices. Apart from the THPI/NHPI data, the other two sources of data were manipulated in the statistical computing environment R using various software packages and statistical functions, as well as in Microsoft's Excel software. The datasets were cleaned of problematic properties such as missing observations or typing/input errors, and the

resulting data were coerced to conform with a format conducive for regression analysis. The data are spread across 8 panels (cities) with a time period for each ranging from 173 to 193 months, resulting in an unbalanced panel set.

i. Teranet House Price Index

The Teranet House Price Index is a joint initiative run by Teranet Inc. and The National Bank of Canada, where the index is calculated over averaging "sales pairs" of homes; homes that have been sold at least twice have the rate at which their property values increase or decrease in those two time periods. Only the data pertaining to the scale of CMA are available to the public, however commercial enterprises can obtain data disaggregated down to the Forward Station Area (FSA) scale, and other finer scales. The data availability from the THPI differs from one CMA to another, but most of the series data are available from July of 1990, with Calgary and Edmonton being two of the few to have a later start date (being March of 1999). This index's rate of change at any time period *can* consist of homes that have not been sold within that time period. In such cases, the value change is *inferred* by other homes that have been sold in this period. This allows for homes that have not been resold in a certain time frame to have their estimated value, as a part of the average of that geographic area, imputed. This method is not without significant caveats to those who wish to conduct an analysis using these data. Some of these caveats are even explicitly stated on THPI's website, mainly with regard to endogenous factors not being captured in their method (Teranet Inc. & National Bank of Canada, 2020):

- The dwelling's property type can change between periods
- Arm's-length sales are not captured
- High turn-over frequency can skew data

Additionally, this method relies on an assumption that is explicitly stated; that there is a "constant level quality between the sales in a linear fashion". This can further add to the problem of endogeneity as dwellings often can go through renovations, suffer damage, or are not maintained to resist the daily wear-and-tear on the dwelling.

These caveats have been echoed with regard to the "repeat-sales" approach taken up by some economists, as explored by Li et al. (2006), in their analysis of constructing a hedonic model with data concerning several boroughs in the Ottawa-Gatineau area. In addition to these caveats, another inconvenient challenge in using the data collected for these dwellings is that only homes that have been sold *at least twice* are captured in the data collection. This automatically presents a selection bias when performing any kind of regression analysis; any homes that have not been sold at least twice, let alone those that have not been sold at all, are not captured (though their value is inferred), which can cause significant bias in estimating the necessary coefficients. This adds further to the existing selection bias that is present in this analysis; only Census Metropolitan Areas are considered. Consequently, rural property values, or towns/cities that do not qualify as Census Metropolitan Areas as defined by Statistics Canada, cannot be captured and used in the analysis. However, if there *were* to be some data present from these geographical areas, one could compensate for the selection bias by performing a Heckman Correction; this correction codifies variables in the dataset that have some data, particularly covariate data (whether exogenous or endogenous), but no data for the dependent variable, by means of a 'dummy' variable. These observations would then be weighted by the Inverse-Mills

ratio and have the observations that *do* have all needed data observations, would have their weights be incorporated into a weighted-least-squares (WLS) regression. (Heckman, 1976) Unfortunately, this is not the case here, as this analysis will incorporate needed time-varying covariates to control for omitted variable bias, and as such, not having this data available for those cities/towns/rural areas would only contribute to time-varying omitted variables being a part of the error term that traditional panel data regression methods like Fixed Effects (Within) or Random-Effects models cannot account for. Regardless, this paper's regression methods will incorporate the THPI data acting as the dependent variable, alongside comparing the NHPI data, to see if one fits our models better.

ii. National House Price Index

The National House Price Index (NHPI) is also available at a monthly frequency. The NHPI captures contractors' sales on dwellings that have specific details pertaining to the dwelling remain constant between two consecutive months. In a notable difference from the data collected by the THPI, the NHPI collects this same data about the land on which property lies, while collecting data on three different types of dwellings (single homes, semi-detached homes, and townhouses) available to researchers to distinguish between when selecting their data. As well, some components of the NHPI, according to Statistics Canada, are incorporated into the Consumer Price Index as well, which is also available monthly. In another difference from the THPI, the NHPI accounts for quality changes over time. The survey conducted by Statistics Canada contains a questionnaire that includes detailed questions pertaining to the characteristics of the land and dwelling. Accounting for these quality changes in dwellings and land allows the NHPI to report only the "pure price change over time". As will be covered later, the THPI fits our model better,

providing a higher adjusted R-Squared, more coefficients with signs that are desired, and coefficients with smaller standard errors, though issues with both models will be discussed.

iii. Pollutant and Climate Data

The crux of this analysis relies on the *a priori* assumption that air pollution is an endogenous covariate, and other environmental data, namely wind direction and wind speed, are relevant and strong instruments to combat this endogeneity. Thankfully, all three of these variables are collected at a very high frequency by Environment Canada, going back sometimes decades for *many* different geographic areas, including rural and urban areas that span nearly the entirety of Canada. Data concerning air particulate matter, including particulate matter at the 2.5-micron size fraction ($PM_{2.5}$), are collected by the National Air Pollution and Surveillance Program (NAPS), a division of Environmental and Climate Change Canada (ECCC). The ECCC is facilitated by a multi-provincial and federal partnership in the collection and monitoring of many airborne pollutants. The unit of measurement for the concentration of PM2.5 is $\mu g/m^3$ (i.e. micrograms per cubic metre of air). There are nearly 260 air-quality monitoring stations in over 150 rural and urban geographic locations across Canada, with data spanning back decades, depending on the location. These data are collected every hour of every day of the year, with the results usually posted at a quarterly basis. For the purposes of this analysis, the stations matching a certain CMA are simply the stations that are located within, or just outside of the CMA, to include as much data as possible. The results are then averaged across all these stations, and then averaged to a monthly value for that CMA (a similar method is followed for the wind data).

Previous methods in combining data often use the inverse-weighted or inverse-squared-weighted distance from each station to the relevant geographic area's "center"; this would give stations closest to the geographic center of an area the most weight, and stations further away would have their weights reduced (dramatically so, in the case of inverse-squaring the distance). However, data for a station close to the center of a geographic area might have missing observations or reflect values that are found to often be unique to that station alone, when other surrounding stations less weight might have more "correct" observations. Therefore, in order to not give any undue favoritism to any station in the area, the different stations share equal weight in determining that CMA's monthly pollution level. This same method applies to the wind-direction and wind-speed data. Data pertaining to wind direction and wind speed are collected by the Meteorological Service of Canada (MSC), where wind direction is measured in tens of degrees from the direction in which the wind blows. For an example, a value of 9 means 90 degrees East, or an Eastern wind. A value of zero (0) denotes a calm wind, and a value of 36 indicates 360 degrees, or a wind blowing from the North pole. All these directions are measured in the context of the *geographic* direction, not magnetic direction. Wind speed is denoted in kilometres-per-hour (km/h) measured at a height of 10 metres above the ground.

iv. Covariate Data on Other Control Variables

With our dependent, endogenous covariate, and instrumental variables covered, what remains is the other (assumed) exogenous covariates that are time-varying corresponding to the relevant CMA. These all originate from Statistics Canada's CANSIM data tables. They include:

- Homicide Rate (homicides per 100,000 persons)
- Median After-Tax Household Income (measured/divided by $1,000)
- Low-Income-Cut-Off (LICO) (Percentage of individuals living below the cutoff)
- The Unemployment Rate
- Average Rent of Dwellings
- The vacancy rate of dwellings
- Population (measured/divided by 1,000)

While this analysis follows much of the similar construction of Hong's 2015 hedonic model, Hong included only 3 of the preceding covariates (income, population, and the unemployment rate), with no additional covariates provided other than the endogenous pollution regressor. A noted concern at this time is to control for more time-varying covariates to mitigate the possibility of unobserved time-varying variables biasing the estimation of our coefficients. All these covariates are measured at an annual frequency but are disaggregated via the Denton-Cholette algorithm in the **tempdisagg** package in R to monthly observations.

A brief overview of the desired sign and relevance of each covariate is discussed below with respect to property values.

Homicide rate's desired sign is negative, for what should hopefully be obvious reasons. Troy and Grove (2008) conduct an analysis to determine whether public parks are a desired public amenity, or if there are factors that could determine it to be more of a liability, reflected through a hedonic model capturing proximity to a public park as an exogenous variable, and crime rate reported in various neighbourhoods in Baltimore, Maryland. They find that beyond a certain threshold, property values indeed decline in response to higher-than-tolerable crime rates, and that certain public amenities do not outweigh this burden.

Median after-tax household income's desired sign is positive but this is not easily established in the literature. Families with higher average incomes can afford to purchase more expensive homes that often come with more desirable private and public amenities. This might be counteracted, however, by a seemingly large household after-tax income being made up of a larger-than-average household size, some of whom might move out from their respective dwellings at any time, or who may prefer to pay less for more space to accommodate the larger families, as noted by Bajari and Kahn (2007).

Per Statistics Canada, the Low-Income-Cutoffs (LICO) are thresholds which families or households "will likely devote a larger share of its income on the necessities of food, shelter, and clothing than the average family". The expected sign for the LICO coefficient should be negative. If a larger share of households or persons were to live below these thresholds, they likely would not be able to afford dwellings in pricier areas or cities. As such, they would likely self-select into neighbourhoods/cities where average house prices are lower, to accommodate their higher-than average marginal spending on life necessities. The causality, or determining the extent to whether the "chicken or the egg" in this scenario is dominant with regards to the LICO vis-a-vis house prices, is beyond the scope of this paper, however it seems logical *a priori* that neighbourhood, or city-level, poverty would in some way be correlated with house prices in a negative fashion. In this paper, to eliminate any effect that taxes have on distortions of behaviour, the after-tax measure of LICO is used, as usually after taxes, transfer payments and such have been incorporated, the income gap is compressed, allowing for a more accurate "take-home" measure of income able to affect behaviour. This is supported by research like Leonard et al. (2017), where they find neighbourhood/block-level poverty levels are negatively linked

to other neighbourhood conditions and amenities, by establishing this link down to neighbourhood-level data.

The expected sign for vacancy rate of properties is negative. If vacancy rates are low, one would assume that dwellings for that particular area/city are in sufficiently high demand, and meet a sufficient market-clearing rate, that enough home-buyers find homes at a rate keeping the proportion of those that *do* not find homes (i.e. homes that are vacant, otherwise) far lower. In other words, a high vacancy rate can indicate an excess of supply of dwellings for sale, and *ceteris paribus*, would lead to lower-than-average dwelling prices. Coulson and Zabel (2013) use vacancy rate as an implicit measure of foreclosures. They found that house prices can exhibit "downward sticky" behaviour, as the traditional mechanisms by which hedonic models operate can be biased if the market is dominated by excessive foreclosure, and thereby can display higher vacancy rates. Therefore, because of how vacancy rate can have substantial bias in estimating the coefficients, they also recommend controlling for their effects whenever a hedonic regression model is used.

IV. Methodology and Model Construction

In this paper, the traditional methods of carrying out panel regressions will be included to determine what is the best fit for our model, including the dimension of instrumental variables complicating the construction of each method. The following model construction will be used conduct the analysis in this paper:

$$Price_{it} = \beta_0 + \theta \widehat{Pollution}_{it} + \beta X_{it} + \eta_i + \varepsilon_{it}$$

where: β_0 is our time and panel-invariant constant, which will be included in our typical pooled-OLS/IV or random-effects (RE/REIV) models and dropped in our within (FE/FEIV) models; β is a 1 x K vector containing the coefficients corresponding with our exogenous regressors; X_{it} is a N x K matrix containing exogenous variable data on our observations; $\widehat{Pollution}_{it}$ is the fitted values of pollution by fitting our exogenous and instrumental regressors onto our endogenous Pollution data in the first-stage of our two-stage-least-squares (2SLS) regression, which will be detailed below; θ is the coefficient with regards to our fitted-values Pollution data; η_i is our time-invariant unobserved effect, which will be theoretically eliminated in our Within/Fixed-Effects (FE/FEIV) model; and ε_{it} is our idiosyncratic/time-varying error.

The following will be the first-stage regression in our 2SLS method explained above to obtain our fitted values of our endogenous Pollution regressor:

$$Pollution_{it} = \alpha_0 + \alpha Z_{it} + \upsilon_i + \omega_{it}$$

As established before, Z_{it} is a N x L (where L ≥ K) matrix containing our exogenous and instrumental regressors; α is a 1 x L vector containing the coefficients for our exogenous regressors; α_0 is our constant that, like our second-stage regression, will be included in only our pooled-OLS and random-effects (RE/REIV) models; υ_i is our time-invariant error

for our first-stage regression, which will also be included only in the pooled-OLS and random-effects models; and ω_{it} is our time-varying/idiosyncratic error for the first-stage regression.

The next section will discuss the results of the various models tested, including an examination by which we overcame the alleged endogeneity in our Pollution regressor.

V. Model Results and Diagnostics

i. Non-IV Estimation Benchmarks

To begin this section, we will present the results of a pooled-OLS model as our benchmark. No random or fixed-effects estimation is utilized. The results are displayed below:

Table 1: THPI vs NHPI OLS Results

	Dependent variable:	
	THPI	NHPI
	(1)	(2)
Pollution	1.480***	0.537***
	(0.244)	(0.127)
Homicide Rate	4.938***	-1.350***
	(0.665)	(0.347)
Median Income	-3.232***	-0.999***
	(0.140)	(0.073)
LICO	-6.859***	-2.046***
	(0.317)	(0.166)
Rent	0.208***	0.074***
	(0.005)	(0.003)
Unemployment Rate	1.134*	1.668***
	(0.643)	(0.336)
Vacancy Rate	1.700***	0.828***
	(0.500)	(0.261)
Population (per 1,000)	-0.007***	-0.004***
	(0.001)	(0.0003)
Constant	209.617***	97.117***
	(8.915)	(4.659)
Observations	1,507	1,507
R^2	0.643	0.459
Adjusted R^2	0.642	0.456
F Statistic (df = 8; 1498)	337.904***	159.076***
Note:		*p<0.1; **p<0.05; ***p<0.01

As is clear here, and in the following tables, the THPI gives a better fit to our model, yielding a higher adjusted-R^2, however at the cost of an undesirable sign (positive) for the

Homicide regressor. The value of the Pollution regressor is very close to zero in the NHPI model, and for the THPI model, gives us the undesirable positive sign. In addition, the Vacancy Rate regressor, believed *a priori* to be negatively correlated with both price indices, has however the opposite correlation in both models. The following table introduces random-effects to our OLS estimation to account for individual-specific effects that may occur.

Table 2: THPI vs NHPI Random-Effects Models

	Dependent variable:	
	THPI	NHPI
	(1)	(2)
Pollution	0.046	0.070
	(0.127)	(0.074)
Homicide Rate	-0.080	2.642***
	(0.562)	(0.329)
Median Income	-0.998***	0.015
	(0.106)	(0.062)
LICO	-2.039***	-0.145
	(0.188)	(0.110)
Rent	0.262***	0.101***
	(0.005)	(0.003)
Unemployment Rate	-6.295***	-2.342***
	(0.334)	(0.196)
Vacancy Rate	2.299***	1.003***
	(0.259)	(0.152)
Population (per 1,000)	0.028***	0.008***
	(0.002)	(0.001)
Constant	-29.945	-10.662
	(26.598)	(10.503)
Observations	1,507	1,507
R^2	0.912	0.820
Adjusted R^2	0.911	0.819
F Statistic	15,492.560***	6,810.869***
Note:		*p<0.1; **p<0.05; ***p<0.01

After introducing random effects to our model, the Pollution regressor's value hovers close to zero in both models, the sign for the Homicide regressor is reversed in both models, and

our R^2 is substantially higher, and the THPI model still performs better than the NHPI model. After running both models in the "within" case to introduce individual fixed effects, and in the random-effects scenarios, a Hausman test was conducted to see whether one model is more consistent. Arriving at a χ_8^2 value of 6.4053, this results in a p-value of 0.6019. Thus we do not reject the null hypothesis, concluding that although both models are consistent, the random-effects models is the more efficient. Likewise, to determine if there any significant individual-specific effects present in the dataset, we conducted a Breusch-Pagan Lagrange Multiplier test, arriving at a χ_1^2 value of 8835. In this case, the resulting p-value is far below 0.001, thereby soundly rejecting the null hypothesis that there are no significant effects. Alas, we are justified introducing individual-specific effects into our non-IV models.

Though the fit of our models improves in both the THPI and NHPI cases when individual-specific effects are introduced, we find an undesirable and insignificant magnitude for our Pollution regressor. To seek a more believable model, a model more representative of our reality, we will introduce Wind Direction and Wind Speed instruments to represent the Pollution regressor seeking more useful results.

ii. Instrumental Variable Estimation

In addressing the endogeneity of our Pollution regressor, Wind Direction and Wind Speed are used as instrumental variables in both a pooled-OLS and random-effects model. The following table uses these two instruments to fit values for our Pollution regressor using 2SLS to address its endogeneity. The results are displayed in Table 3:

Table 3: Pooling vs Random-Effects IV Estimation

	Dependent variable:			
	THPI		NHPI	
	(1)	(2)	(3)	(4)
Pollution	-5.081**	-8.104***	-6.895***	-3.464***
	(2.289)	(1.574)	(1.777)	(0.755)
Homicide Rate	2.486**	-3.444***	-4.128***	1.186**
	(1.173)	(1.253)	(0.910)	(0.600)
Median Income	-2.998***	-1.212***	-0.733***	-0.070
	(0.189)	(0.208)	(0.147)	(0.100)
LICO	-8.197***	-1.389***	-3.561***	0.138
	(0.603)	(0.386)	(0.468)	(0.185)
Rent	0.180***	0.317***	0.043***	0.124***
	(0.011)	(0.014)	(0.009)	(0.006)
Unemployment Rate	2.623***	-5.122***	3.355***	-1.836***
	(0.938)	(0.684)	(0.728)	(0.328)
Vacancy Rate	2.632***	1.708***	1.884***	0.760***
	(0.689)	(0.513)	(0.535)	(0.246)
Population (per 1,000)	-0.004***	0.003	-0.00002	-0.002
	(0.001)	(0.006)	(0.001)	(0.003)
Constant	269.516***	41.769	164.978***	19.091
	(23.395)	(25.617)	(18.161)	(12.101)
Observations	1,507	1,507	1,507	1,507
R^2	0.495	0.718	0.091	0.612
Adjusted R^2	0.492	0.716	0.086	0.610
F Statistic	1,801.790***	4,118.936***	398.269***	2,720.917***
Note:				*p<0.1; **p<0.05; ***p<0.01

For Table 3, the first and third columns are the pooled-IV models, while the second and fourth columns are our random-effects IV models. Here we see a stark contrast to the situation before with our benchmark OLS estimation. The most significant and helpful change is that our Pollution regressor has reversed signs from positive to negative, signifying that a more powerful representation of observed realities has been created in the THPI random-effects IV model. Additionally, all cases present a substantial magnitude of an impact from our Pollution regressor, with the NHPI RE/IV model showing the weakest

magnitude of -3.464. Our pooled models show the weakest fits, with the NHPI model showing only an adjusted-R^2 value of 8.6%. This is in significant contrast with both of the RE/IV models, both having adjusted-R^2 values exceeding 60%, again with the THPI model showing the best fit, and the highest number of regressors having desirable signs. Both our Homicide Rate and Unemployment Rate regressors have desirable negative signs in the THPI RE/IV model, unlike results from two of the other models implemented. However, Median Household Income and Vacancy Rate have undesirable signs across all models implemented. This could because of several issues which were discussed in the proceeding section.

As before, we find the random-effects to provide the best fit compared with that of fixed-effects and pooled-IV implementation; conducting a Hausman test we arrive at a χ_8^2 of 1.9713, with its corresponding p-value of 0.4831. We do not reject the null, and find the random-effects model more efficient than a fixed-effects model. Conducting a Breusch-Pagan LM test we arrive at a χ_1^2 value of 7950.6 and a corresponding p-value of well below 0.001. Thus we reject the null and find statistically significant individual-specific effects in our data, even when accounting for regressor endogeneity, thereby finding our random-effects model to be a better performer than our pooled-IV model.

Examining whether quadratic regressors should be included in the model, Ramsey's RESET test for functional form was conducted, producing a χ_1^2 value of 0.4919 having a p-value of 0.4831. We therefore fail to reject the null, and conclude that adding the squared fitted values from our RE/IV model does not add sufficient significant effects to our model to warrant including quadratic regressors.

iii. Validity of Instruments

In using two instrumental variables to represent one alleged endogenous regressor, additional to testing whether these instruments are strong enough to warrant their use in an IV model and whether Pollution is an endogenous regressor, overfitting of the model must be examined to see whether multiple instruments are needed to supplant the Pollution regressor. This is done using the Sargan test, later in this section. And because this is not a single cross-section worth of data making up the IV model, the first-stage regression must incorporate the random-effects transformation that we apply to the 2SLS model as well. Because the THPI gave a better fit compared with the NHPI model in terms of adjusted-R^2 and the most number of regressors with desirable signs, we will examine only the THPI model going forward. The following table gives the first-stage regression results of regressing Pollution against all exogenous regressors and the two instruments:

Table 4: First Stage Random-Effects Estimation

	Dependent variable:
	Pollution
Wind Direction	0.066^{***}
	(0.021)
Wind Speed	-0.053^{***}
	(0.010)
Constant	9.097^{***}
	(2.701)
Observations	1,507
R^2	0.072
Adjusted R^2	0.067
F Statistic	116.987^{***}

Note: All other covaries are controlled for but are omitted in this table.

$^{*}p<0.1$; $^{**}p<0.05$; $^{***}p<0.01$

While our overall fit of the model is not strong, both instrumental variables are statistically significant at the 1% level. In addition, a Wald test comparing the first-stage model with

the instruments omitted yields a χ_2^2 value of 36.782 having a p-value well below 0.001. We thus reject the null and find these two instruments are strong and relevant for the model.

The connection between concentration of various pollutants or 'dust' and several meteorological characteristics of an area are well documented in the literature; Csavina et al. (2014) extract a complex non-linear relationship among meteorological data for wind direction, wind speed, relative humidity, and local PM_{10} concentrations in several locations in the Southwest United States. A similar study conducted by Wang and Ogawa (2015) on the relationship among wind direction, wind speed, and $PM_{2.5}$ concentrations in and surrounding Nagasaki, Japan found that above or below a certain threshold of wind speed, the correlation to $PM_{2.5}$ concentration could either be positive or negative. What proves advantageous for our analysis is that those analyses usually found that above a certain threshold value for wind speed (above 3m/s, per Wang and Ogawa), the correlation between wind speed and $PM_{2.5}$ concentrations is *negative*. This seems to confirm the validity of the sign derived for our Wind Speed regressor (Wind Speed has a minimum value of 9.57 km/h in our analysis, substantially above the Wang and Ogawa 3 m/s threshold) in our first-stage regression. In terms of meteorological intuition, higher wind speeds should logically dilute and disseminate local particulate, thereby reducing concentrations in an area of interest. This suggest the necessity to use this information to better account for the variation in house price indices when local pollutant concentrations change due to meteorological conditions in a CMA of interest.

As stated earlier, what remains now is to determine whether our model is overfitted by including two instruments for one endogenous variable, instead of using only one, and leaving our model exactly identified in moment conditions. We conduct the

Sargan test and arrive at χ_1^2 value of 15.106, soundly rejecting the null that our model is justly identified. Because we have found that our model is overidentified in its restrictions, it must be determined which is the better instrument, Wind Direction or Wind Speed.

Table 5: Comparison of Single-Instrument IV Regressions

	Dependent variable:	
	THPI	
	(1)	(2)
Pollution	4.308*	-11.494***
	(2.408)	(2.432)
Homicide Rate	1.645	-4.825***
	(1.225)	(1.721)
Median Household Income	-0.895***	-1.286***
	(0.152)	(0.274)
LICO	-2.394***	-1.133**
	(0.319)	(0.519)
Rent	0.235***	0.337***
	(0.017)	(0.019)
Unemployment Rate	-6.891***	-4.627***
	(0.556)	(0.918)
Vacancy Rate	2.598***	1.492**
	(0.383)	(0.679)
Population	0.040***	-0.006
	(0.008)	(0.008)
Constant	-65.571*	69.419**
	(38.352)	(31.227)
Observations	1,507	1,507
R^2	0.850	0.591
Adjusted R^2	0.849	0.589
F Statistic	8,825.667***	2,372.361***
Note:		*p<0.1; **p<0.05; ***p<0.01

When reading Table 5, where Wind Direction is the only instrument used in the model on the left, while Wind Speed is the only instrument used in the model on the right, certain insights emerge in the table. First, according to the literature, Wind Direction was found to account better for the endogeneity of the Pollution regressor, yet here Wind Speed extracts the desired sign for the endogenous Pollution regressor, and at considerable magnitude (-11.494), while being more statistically significant than Wind Direction, even though

Pollution through Wind Direction is statistically significant as well at the 5% threshold. Second, we derive a desirable sign for the Homicide regressor as well, leaving the second model with the most regressors having the most desirable signs of the two, despite losing a considerable amount of the model's explanatory power (i.e. a lower adjusted-R^2). This is disappointing as Bondy et al. (2018) found that Wind Direction was the primary means by which they could account for the endogeneity in their Pollution regressor when examining London inner-city boroughs. This might be an artifact arising from the various levels of temporal (dis)aggregation required to create our dataset for this analysis, or perhaps it presents an alternate explanatory framework more descriptive of the meteorology within various Canadian metropolitan areas. In the Canadian urban context, Wind Speed is found to be a more significant factor to control for when examining the problem of incorporating Pollution into hedonic regression models. Regardless, though we have achieved the desired sign for our Pollution regressor with a decent fit, there remain several problems with the model which will be discussed in the following section, particularly various correlation issues.

iv. Model Caveats

In the preceding section, we found a desirable random-effects instrumental variable (REIV) model captured well the effects of Pollution on the Teranet House Price Index when controlling for various exogenous covariates and endogeneity by instrumenting the Pollution regressor with the Wind Speed variable. However, two forms of correlation are salient issues when discussing the inferential strength of the model established. Firstly, cross-sectional correlation: conducting Pesaran's Cross-Sectional Dependence Test for unbalanced panels yields a Z value of 16.06, with a corresponding p-value of well below

0.001, thus rejecting the null hypothesis of no cross-sectional correlation between groups. However, this result should not be too surprising, as our various cross-sections (Canadian Metropolitan Areas) are not randomly drawn for this analysis, and the changes of one city can plausibly affect the circumstances/characteristics of another city, which further reinforces the selection bias described earlier. Accordingly, including more municipalities/town with all their available data pertaining to our exogenous/endogenous and instrumental variables, while potentially missing price data from the THPI or the NHPI, would be a prudent follow-up examination. Conducting a Heckman correction for sample-selection biases, one could account for the biases occurring from the non-random samples we have here.

The second issue is serial autocorrelation in the errors, to which a Dynamic Panel Data model can be applied to correct for this issue, as in Arellano & Bond (1991) who use a Dynamic Generalized-Method-of-Moments (GMM) estimation. While serial autocorrelation will not bias our estimators, it will bias their variances. Conducting a Breusch-Godfrey test for serial correlation yields a χ^2_{178} value of 1477.7, suggesting that we reject the null hypothesis of no serial correlation in the errors of our REIV model. Due to the attributes of the data used in this paper, our panel has a small number of geographical cross-sections (only 8 CMA's), and a large number of months for which these cross-sections have data available or imputed for (from 178 to 193 months). The dataset could therefore be described as a "small-N, large-T" dataset. Often, the desirable characteristics in which Dynamic Panel Data modeling is conducted are the opposite, where there are many cross-sections available, and the length of time available to each cross-section is small. This is due to having a large headroom for incorporating many regressors and

instruments, which are often time-lagged (allowing also the inclusion of lagged dependent variables) by various lengths to capture the effects of past values of the regressors affecting the value of the dependent variable at time T = t. Because our dataset only has 8 cross-sections, we do not have much headroom for including lagged regressors/dependent variables before we overfit our model and estimation becomes highly constrained, perhaps even not possible (i.e. K > N will occur very quickly). As such, in order to not estimate a Dynamic Panel model with very few regressors, we will correct the standard errors of our estimators by utilizing the covariance matrix estimation proposed by Driscoll and Kraay (1998) which was found to be robust against cross-sectional *and* serial correlation in a T-asymptotic context regardless of the dimension of N, which describes our large-T dataset aptly. Upon correcting for these two biases we are left with the following coefficients and their standard errors:

Table 6: Robust Estimation of Coefficients

	THPI
Pollution	-11.494***
	(2.109)
Homicide Rate	-4.825**
	(2.053)
Median Household Income	-1.286***
	(0.262)
LICO	-1.133
	(0.747)
Rent	0.337***
	(0.019)
Unemployment Rate	-4.627***
	(1.568)
Vacancy Rate	1.492
	(1.360)
Population	-0.006
	(0.008)
Constant	69.419*
	(41.345)
Note:	*p<0.1; **p<0.05; ***p<0.01

We lose statistical significance on some of our regressors compared with our homoscedastic-REIV model, however our Pollution regressor is still statistically significant, which again is instrumented via Wind Speed.

In summary, though some issues of serial and cross-sectional correlation remain present, correcting for these biases in the standard errors in our REIV model still leaves us with the desired sign for our Pollution regressor, which is highly statistically significant. The magnitude of our Pollution regressor infers that an increase of $PM_{2.5}$ concentration ($\mu g/m^3$) either from one CMA to another, or from one month to another within the year, will yield, *ceteris paribus,* an average decrease of ≈ 11.5 points in the Teranet House Price Index.

VI. Conclusion

Though the hedonic pricing model is a common method of measuring the impacts of various pollutants on real estate markets, regressor endogeneity is a recurring issue in many analyses. As such, its influence should be counteracted by implementing proper instrumental variables in the hedonic model used. Previous literature established Wind Direction *a priori* to be the better instrumental variable for representing the Pollution regressor. However, the results of this study suggest that in Canadian Metropolitan Areas, Wind Speed is a superior instrument which both yields the desired negative sign for the Pollution regressor, and results in the highest number of regressors displaying desired signs. There may be an issue in this approach regarding performance of temporal disaggregation *and* aggregation on the various data sources within the subject dataset, whereby some of the effect of a certain variable is diminished or absorbed by being aggregated to a lower frequency, or disaggregated to a higher frequency. Regarding temporal *dis*aggregation, using a proper indicator time-series set would be an improved method for future research; the variable being temporally disaggregated can better impute its values at that higher frequency.

Additionally, this analysis contains both a selection bias and a cross-sectional dependence bias, due to the nature of the sample selected (Canadian metropolitan areas) not being random. Two steps may be taken to correct this: the first is to include all municipalities and cities that have data recorded for exogenous variables. The second is to conduct a Heckman Correction on those that *do not* have data for the dependent variable, which would help correct for the observations that *do* have data on the dependent variable.

Finally, researchers require access to more pricing data for the Canadian housing market, metropolitan or otherwise, in order to conduct a more thorough analysis of this subject. Access to these data would undoubtedly shed further light on air pollution's true impact on Canada's real estate market. Because of the toxicity of suspended particulate matter, especially $PM_{2.5}$ (CEPA 2007), the ability to use these data could also allow exploration of the urban regional public-health implications of TSP, especially $PM_{2.5}$, and whether areal concentration distributions can predict real estate price indices.

References

Anderson, M. (2015). As The Wind Blows: The Effects of Long-Term Exposure to Air Pollution on Mortality. *National Bureau of Economic Research*.

Arellano, M., & Bond, S. (1991). Some tests of specification for panel data: Monte Carlo evidence and an application to employment equations. *Review of Economic Studies, 58*(2), 277-297.

Bajari, P., & Khan, M. (2007). Estimating Hedonic Models of Consumer Demand with an Application to Urban Sprawl. In A. Baranzini, J. Ramirez, C. Schaerer, & P. Thalmann, *Hedonic Methods in Housing Markets* (pp. 129-155). New York: Springer.

Bondy, M., Roth, S., & Sager, L. (2018, April). *Crime is in the Air: The Contemporaneous Relationship Between Air Pollution and Crime.* From IZA Institute of Labor Economics: http://ftp.iza.org/dp11492.pdf

Canadian Environmental Protection Act (CEPA). 2007. *Priority Substances List Assessment Report for Respirable Particulate Matter.* Available at: www.canada.ca › reports-publications › environmental-contaminants.

Chan, W. M. (2014, July). *Comparison of Spatial Hedonic House Price Models: Application to Real Estate Transactions in Vancouver West.* Retrieved March 14, 2020 from Simon Fraser University: http://summit.sfu.ca/system/files/iritems1/14416/FINAL%20PROJECT%20Wai%20Man%20Chan.pdf

Chay, K., & Greenstone, M. (2005). Does Air Quality Matter? Evidence from the Housing Market. *Journal of Political Economy, 113*(2). Retrieved February 15, 2019

Cholette, P.-A., & Dagum, E. B. (2006). Benchmarking, Temporal Distribution, and Reconciliation Methods for Time Series. In *Lecture Notes in Statistics* (pp. 80,82). New York: Springer-Verlag.

Chow, G., & Lin, A. (1971). Best linear unbiased interpretation , distribution, and extrapolation of time series by related series. *The Review of Economics and Statistics, 53*(4), 372-375.

Coulson, N. E., & Zabel, J. E. (2013). What Can We Learn from Hedonic Models When Housing Markets Are Dominated by Foreclosures? *Annual Review of Resource Economics, 5*(1), 261-279.

Croissant, Y., & Millo, G. (2008). Panel Data Econometrics with R: the plm package. *Journal of Statistical Software, 27*(2), 1-43. doi:10.18637/jss.v027.i02

Csavina, J., Field, J., Felix, O., Corral-Avita, A. Y., Saez, A. E., & Betterton, E. A. (2014, July 15). Effect of Wind Speed and Relative Humidity on Atmospheric

Dust Concentrations in Semi-Arid Climates. *Science of the Total Environment, 487*, 82-90. doi:10.1016/j.scitotenv.2014.03.138

Denton, F. T. (1971). Adjustment of Monthly or Quarterly Series to Annual Totals: An Approach Based on Quadratic Minimization. *Journal of the American Statistical Association, 66*(333), 99-102.

Deryugina, T., Heutel, G., Miller, N., & Reif, J. (2016). The mortality and medical costs of air pollution: Evidence from changes in wind direction. *National Bureau of Economic Research*.

Driscoll, J., & Kraay, A. (1998). Consistent Covariance Matrix Estimation With Spatially Dependent Panel Data. *The Review of Economics and Statistics, 80*(4), 549-560.

Garrett, T. (2002). Aggregated vs. disaggregated data in regression analysis: Implications for inference. *Federal Reserve Bank of St. Louis, 2002*(024). Retrieved October 20, 2019

Government of Canada. (2019, July 16). *National Air Pollution Surveillance Program*. Retrieved November, 2019 from Government of Canada: https://www.canada.ca/en/environment-climate-change/services/air-pollution/monitoring-networks-data/national-air-pollution-program.html

Government of Canada. (2020, January). *Historical Climate Data*. Retrieved November, 2019 from Government of Canada: https://climate.weather.gc.ca/

Heckman, J. (1976). The Common Structure of Statistical Models of Truncation, Sample Selection and Limited Dependent Variables and a Simple Estimator for Such Models. *Annals of Economic and Social Measurement, 5*(4), 475-492.

Hong, S. (2015, August 26). *Does air pollution feature lower housing price in Canada?* (Y. Aydede, & A. Akbari, Eds.) From Saint Mary's University: http://library2.smu.ca/bitstream/handle/01/26426/Hong_Sheng_MRP_2015.pdf?sequence=1&isAllowed=y

Jacobs, J. (1994). *'Dividing by 4': a feasible quarterly forecasting method?* University of Groningen, Department of Economics.

Leonard, T., Powell-Wiley, T. M., Ayers, C., Murdoch, J. C., Yin, W., & Pruitt, S. L. (2016, July). Property Values as a Measure of Neighbourhoods: An Application of Hedonic Price Theory. *Epidemiology, 27*(4), 518-524.

Li, W., Prud'homme, M., & Yu, K. (2006). Studies in Hedonic Resale Housing Price Indexes. *Canadian Economic Association 40th Annual Meetings.* Montreal: Concordia University.

Litterman, R. (1983). A random walk, Markov model for the distribution of time series. *The Review of Economics and Statistics, 1*(2), 471-478.

Muller-Kademann, C. (2015). Internal Validation of Temporal Disaggregation: A Cloud Chamber Approach. *Journal of Economic and Statistics, 235*(3), 298-319.

Rosen, S. (1974, Jan. - Feb.). Hedonic prices and implicit markets: product differentiation in pure competition. *The Journal of Political Economy, 82*(1), 34-55. Retrieved October 21, 2019

Sax, C., & Steiner, P. (2013, December). Temporal Disaggregation of TIme Series. *The R Journal, 5*(2), 80-87.

Small, K. (1975). Air pollution and property values: further comment. *Review of Economics and Statistics, 57*, 105-107.

Statistics Canada. (2015, November 27). *Low income cut-offs*. From Statistics Canada: https://www150.statcan.gc.ca/n1/pub/75f0002m/2012002/lico-sfr-eng.htm

Teranet Inc. & National Bank of Canada. (2020). *Our Methodology*. From House Price Index: Teranet and National Bank of Canada: https://housepriceindex.ca/about/our-methodology/

Troy, A., & Grove, J. M. (2008). Property values, parks, and crime: a hedonic analysis in Baltimore, MD. *Landscape and Urban Planning, 87*(3), 233-245.

Wang, J., & Ogawa, S. (2015, August 12). Effects of Meteorological Conditions on PM2.5 Concentrations in Nagasaki, Japan. *International Journal of Environmental Research and Public Health, 12*(8), 9089-9101. doi:10.3390/ijerph120809089

Wei, W. W., & Stram, D. O. (1990). Disaggregation of Time Series Models. *Journal of the Royal Statistical Society, 52*(3), 453-467.

Appendix

A. Summary Statistics of Data

Table 7: Summary Statistics of All Data Used

Statistic	N	Mean	St. Dev.	Min	25%	75%	Max
THPI	1,507	132.257	39.111	63.990	96.175	164.730	249.530
Wind Direction	1,507	22.175	3.509	10.182	19.996	24.514	32.400
Wind Speed	1,507	38.695	11.527	9.567	38.817	45.456	59.364
Pollution	1,507	7.289	2.737	2.294	5.355	8.751	22.717
Homicide Rate	1,507	1.941	1.065	-0.220	1.222	2.629	5.427
Median Income	1,507	58.736	8.161	43.810	52.390	62.908	86.519
LICO	1,507	11.623	3.132	4.109	9.145	14.098	19.324
Rent	1,507	839.645	202.039	502.829	678.978	984.696	1,370.965
Unemployment Rate	1,507	6.224	1.493	2.780	5.115	7.265	10.045
Vacancy Rate	1,507	2.275	1.417	0.241	1.247	3.144	7.501
Population	1,507	2,102.551	1,642.983	694.968	962.461	3,528.748	6,300.593
NHPI	1,507	83.229	16.597	43.100	72.150	97.550	116.700

Table 8: Summary Statistics of PM2.5 by CMA

City	Mean	St.Dev	Min	Pctl25	Pctl75	Max
Calgary	7.751759	2.943005	3.609768	5.788439	9.066864	22.71700
Edmonton	7.637009	2.922817	3.532206	5.667863	8.745143	21.76546
Montreal	9.411391	2.472093	5.003084	7.752516	10.945052	20.29551
Ottawa	6.494898	2.021477	2.614247	4.959157	7.395112	14.15540
Quebec	8.467101	2.235136	4.695193	6.707065	9.786245	15.05940
Toronto	7.976211	2.576290	3.173067	6.005008	9.423589	16.55883
Vancouver	4.899058	1.413597	2.294248	3.899672	5.845022	12.32570
Winnipeg	5.437995	1.600464	2.664074	4.323558	6.031791	13.14554

Table 9: Summary Statistics of THPI by CMA

City	Mean	St.Dev	Min	Pctl25	Pctl75	Max
Calgary	140.1132	37.64993	74.90	96.1700	170.1200	188.35
Edmonton	141.0873	40.26699	68.51	95.4925	173.4975	187.91
Montreal	119.7770	30.54824	63.99	94.1900	149.4900	160.47
Ottawa	116.7997	23.12802	71.80	98.5100	140.2800	147.49
Quebec	130.4234	40.02894	65.85	93.3800	173.1100	183.19
Toronto	123.4710	34.14376	75.69	96.2700	147.3300	218.30
Vancouver	145.7239	44.01499	72.97	106.7100	170.7550	249.53
Winnipeg	141.9777	47.67357	68.46	93.5400	189.9500	204.03

B. Figures and Charts

Figure 1: THPI Growth Over Time By CMA

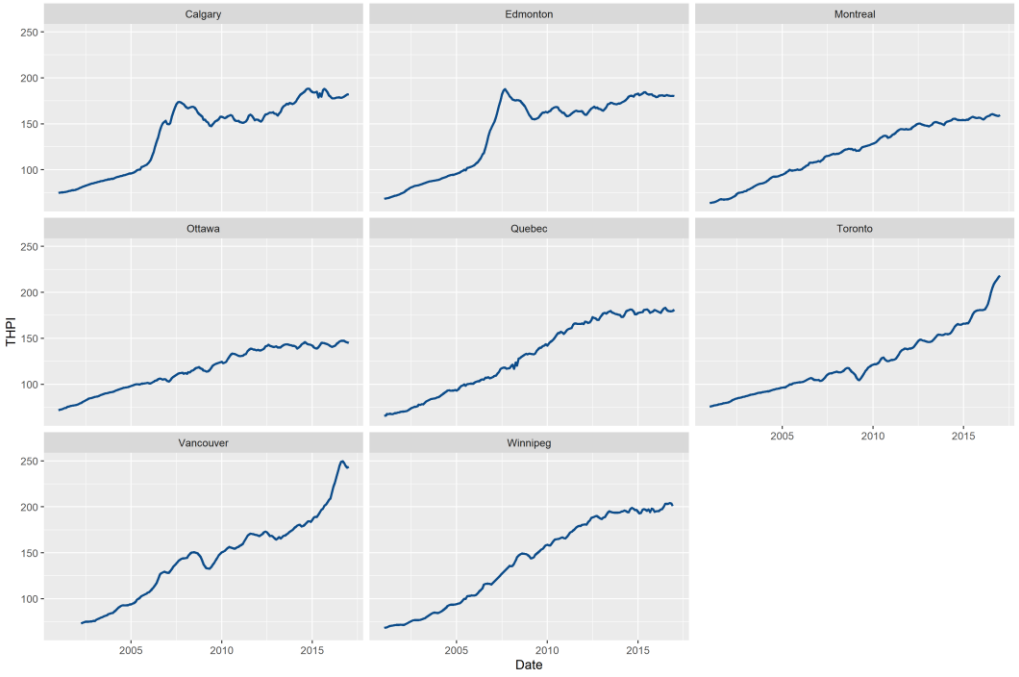

Figure 2: PM2.5 Concentrations Over Time by CMA

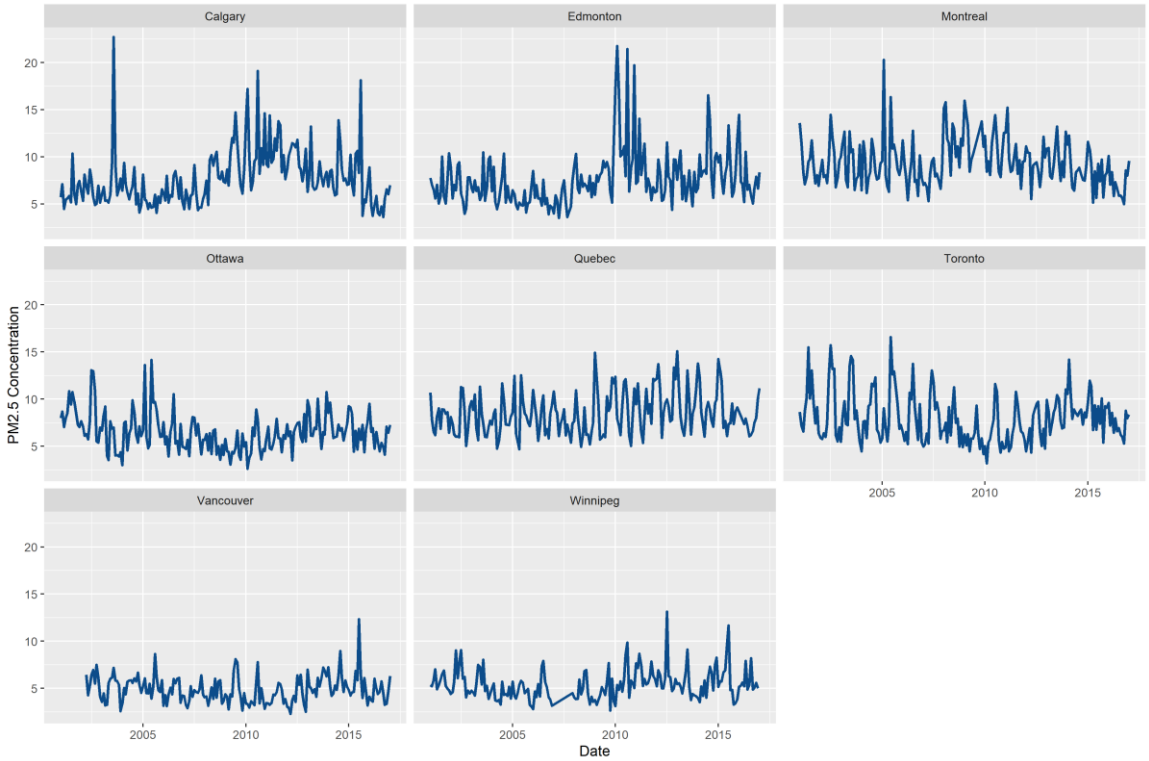

Figure 3: Wind Direction over Time by CMA