**Estimating the Effects of Crime Maps on House Prices using an (Un)natural Experiment: A Study Protocol**

Meng Le Zhang\*[1], Monsuru Adepeju [2], Rhiannon Thomas [3]

[1] Sheffield Methods Institute, University of Sheffield, Sheffield, United Kingdom [2] Department of Sociology, Manchester Metropolitan University, Manchester, United Kingdom [3] Department of Social Science, University of Sheffield, Sheffield, United Kingdom \*Corresponding author: [meng\_le.zhang@sheffield.ac.uk](mailto:meng_le.zhang@sheffield.ac.uk)

# Estimating the Effects of Crime Maps on House Prices using an (Un)natural Experiment: A Study Protocol

# Abstract

Street-level crime maps are publicly available online in England and Wales. However, there was initial resistance to the publication of such fine-grained crime statistics, which can lower house prices and increase insurance premiums in high crime neighbourhoods. Identifying the causal effect of public crime statistics is difficult since crime statistics generally mirror actual crime.

To address this question empirically, we would ideally experiment and introduce a source of random variation in the crime statistics. For instance, we could randomly increase or decrease the number of offences displayed in crime statistics and measure their effects on local house prices. For obvious reasons, we cannot pursue this research design. However, street-level crime maps contain intentional errors, which are the product of a geomasking algorithm designed to mask the location of crimes and protect the identity of victims. This project leverages features associated with the geomasking algorithm to estimate the effect of public crime statistics on house prices.

# Background

Crime may affect house prices through mediating causal pathways–such as the destruction of property or victimisation of locals (Ceccato & Wilhelmsson, 2011; Gibbons, 2004). One mediating pathway is the ‘signalling’ effect of crime which may decrease house prices in high crime areas due to a perception of increased victimisation or other factors like undesirable neighbours. The public may form their opinions about crime from several sources, from word of mouth to official statistics. Since 2011, the UK government has published online monthly crime figures at almost street-level resolution (Sampson and Kinnear 2010). Since launch, the crime map website (henceforth referred to by its domain name police.uk) has received approximately 1 million visits a month (Semrush, 2022). On launch in February 2011, police.uk received over 18 million visits an hour which caused the website to crash repeatedly (Travis and Mulholland 2011). Before the publication of online crime maps on police.uk, street-level crime data was never available to the public (except in West Yorkshire, Sampson and Kinnear 2010).

The release of online crime maps was supposed to create public confidence in crime statistics and support public service transparency (Chainey and Tomspon 2012). The publication of these maps was met by opposition who feared it would affect house prices and increase insurance premiums, particularly in high crime low-income neighbourhoods (Travis and Mulholland 2011, also see Ratcliffe 2002). At the time, the UK government argued that the benefits of open public service data outweighed these concerns (Herbert 2011). However, the advantages and disadvantages of publicly available crime data as a policy are unclear. Many countries with similar capabilities do not make such information public. For example, Scotland has not followed the rest of the UK in this respect.

In this project, we are interested in the causal effect of public crime maps on house prices. This effect is separate from the impact of crime on house prices through other causal pathways such as damage to the environment or negative public perceptions due to victimisation in the local area. The effect of crime maps on house prices is difficult to identify since accurate crime statistics would perfectly mirror actual crime. It is impossible to estimate the effects of crime statistics from actual crime separately in this scenario.

To address this question empirically, we would ideally conduct an experiment and introduce a source of variation in the crime statistics. For instance, across England and Wales, we could randomly increase or decrease the number of offences displayed on police.uk for a period to measure the effects. For obvious reasons, we cannot pursue this research design. However, intentional errors introduced into police.uk crime maps can be leveraged as a source of variation to investigate the signalling effect of the crime maps on the house prices.

Police.uk implements a geomasking algorithm that obscures the actual location of crimes in order to protect the identity of victims. The geomasking algorithm allocates crimes to a nearby geographical location called a snap-point (Tompson et al 2015, Ratcliffe 2002). Whilst the density of snap-points in urban areas is very high, the level of crime in a small area measured using police.uk can differ substantially from actual police records due to geomasking (Tompson et al 2015). This error level gets progressively worse at smaller spatial scales. In 80% of postcodes, the local area crime counts as recorded by police.uk is substantially different from that measured by actual police sources (Tompson et al 2015, figure 7). In short, local area crime statistics on police.uk can be considerably lower or higher than what they should be. This mismatch (henceforth the geomasking error) constitutes the source of variation that we will use in our (un)natural experiment. A key part of this study is that the secret list of snap-points used by police.uk can be inferred using public data alone, as we explain in the next section.

The primary research questions are:

* RQ1. Did police.uk crime statistics affect property prices?
* RQ2. What is the effect of a one-unit increase in crime around a house (as reported by crime maps) on its selling price?

The objective of this study protocol is to specify the research plan ahead of data collection/ access. We can use public domain data to answer RQ1 and we have tried this approach on data from South Yorkshire in the first instance. Our intention is to extend the analysis to all of England and Wales. The early results are in the supplementary materials. RQ2 cannot be answered without access to the original geocoded crime data used by police.uk in their online crime maps. Individual police forces supply this data to the UK Home Office. Our research design also requires access to police force geocoded data before the launch of online crime maps in 2011. To this end, we have gained the support of the South Yorkshire Police (henceforth SYP) Force to access and use their geocoded crimes and incidence data. SYP will also continue to advise us on the processing of the crime data. At this time of writing, we have not yet accessed the police force data needed for RQ2. If data access is granted before this protocol is finalised, we will stop any house price data (i.e. outcomes) from being analysed alongside the police force data. This should be relatively straightforward as we can only access the police data from local machines within SYP settings.

We can use public domain data to answer RQ1. We serendipitously discovered this whilst developing this study protocol. This project and our research design was conceived and funded before we made our serendipitous discovery. RQ1 was part of the original funding application. The hypothesis for RQ1 echoes the position of the government and critics of UK crime maps when it launched in 2011 (i.e. not affected by our early study findings). We have not changed them to reflect our early accidental findings but want to make this clear for readers to judge for themselves. The early results are in the supplementary materials.

# Previous studies

To our knowledge, no other study has studied the effect of crime maps on house prices or leveraged the type of design developed in this study. On the other hand, several observational studies have used crime count data in the UK in the investigation of the relationships between crime and house prices with the results revealing that higher local crime rates negatively impact house prices (Gibbons 2004, Braakman 2017). Globally, Ihlanfeldt and Mayock (2009) reviewed 18 hedonic price studies that have included a measure of neighbourhood crime among the explanatory variables. Generally, crime indicators are negatively associated with house prices but the vast majority of studies do not implement research designs to deal with unobserved characteristics that are associated with both neighbourhood crime and house prices.

The existence of individual police force crime maps before police.uk has been documented in several studies (Sampson and Kinnear 2010, Chainey and Tompson 2008). Quinton (2011) trialled the effects of crime maps as a pilot before the launch of police.uk. The trial involved very aggregate level crime maps, and participants seemed to spend very little time perusing them (average, ~50 seconds). The primary outcome was perceptions of crime. Police.uk’s geomasking algorithm has been studied in several papers (Chainey and Tompson 2012, Smith and Heath 2014, Tompson et al 2015, Quinn et al 2019). Tompson et al’s 2015 paper on how geomasking errors were substantial in lower geographies (e.g. postcodes) greatly influenced this project. Finally, information about the data on police.uk is published on a companion data site: data.police.uk. This site is further documented in Smith and Heath (2014), and past versions are available from internet archives. This project would not have been possible without the excellent documentation in these sources.

The general strategy of leveraging differences between what is publicly seen and what is actually the case is well-established in economics. In labour economics, coarsened information about overall course grades has been used to study the ‘signalling’ effects of education on labour market outcomes (Feng and Graetz 2015). Coarsened public information about Radon has been used to study the effects of Radon on house prices (Pinchbeck et al 2020).

The research team has previously undertaken a pilot study to check the practical feasibility of this approach. We used West Midlands police data which had incomplete records on incidents (e.g. Anti-Social behaviour). We found that, on average, crimes on police.uk are almost 100m away from their actual locations as recorded by the police. No house price data was used in the pilot.

# Hypothesis

**H1.** The hypothesis is that police.uk lowered house prices in high crime areas. This was the position of other property analysts and estate agents in 2011 (Travis and Mulholland 2011). The null hypothesis is that police.uk did not affect house prices. This was the position of the Home Office and some property analysts in 2011 following criticisms of the website.

**H2.** An increase in the number of crimes shown on police.uk’s crime map in an area should lower house prices. This is in line with economic theory that perceptions of high crime make houses less desirable. This was also the position of critics of police.uk in 2011 (Travis and Mulholland 2011). The null hypothesis is that the number of crimes shown on police.uk will not decrease house prices. This was the position of the Home Office.

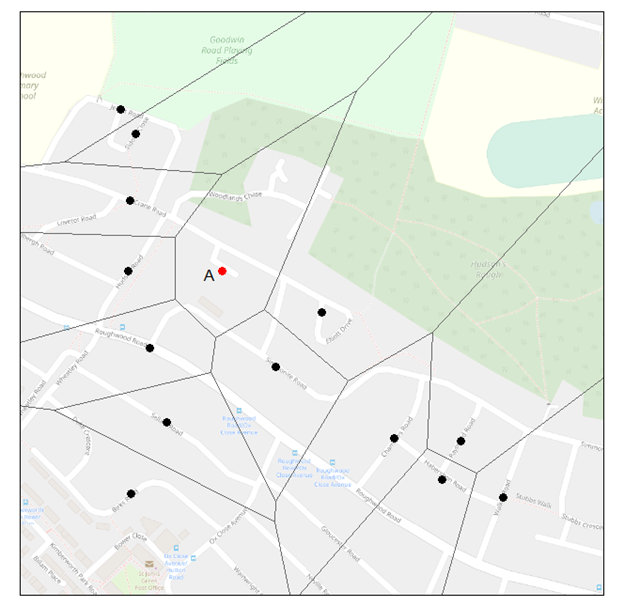
These were the original hypotheses at the onset of the project. Whilst trialing our approach on public domain data for South Yorkshire, we serendipitously uncovered evidence that rejects H1. We have not changed H1 to reflect our knowledge of the data. We have not yet extracted the data required to test H2.

# Geomasking in police.uk

The geomasking routine currently used on police.uk is documented on data.police.uk. Using information from various sources and public data, we know that the geomasking routine was changed in 2012 when police.uk updated its snap-points list. To summarise, the data processing for police.uk is as follows: 1. Home Office receives geocoded data from individual police forces and performs data checks (e.g. identifying crime outside police force boundaries). 2. The coordinates of crimes and incidents are changed to the coordinates of the nearest snap-point for the public data.

We note that police.uk crime maps were updated with a 2-month delay until at least 2016 (as reported by Quinn et al 2019). The list of potential snap points used in step two is a secret. Between 2011 to 2013, we believe there were at least two major versions of the snap-point list which we henceforth refer to as Version One and Version Two. Version One was used on records between Dec 2010 and Nov 2011 whilst Version Two was used on records from Dec 2011 until at least the end of our study period (Dec 2013). The snap points were created in the following ways:

*Version One*: This was created in 2011 using Ordnance Survey data (exact datasets unverified). Potential snap-points were the centre of residential roads (Smith and Heath 2014, p. 93 - 94), and Catchment areas for each point were created using Voronoi polygons. If a snap-point has less than 12 postal addresses within its catchment, it is dropped. This version consists of the remaining snap-points. For computational efficiency, police.uk could have used a nearest neighbour finding algorithm to do the postal address counting. This would have yielded the same outcome as the catchment approach.

 **Figure 1: Example of Voronoi polygons. Polygon A is associated with point A and represents areas closer to point A than any other point. Source: OpenStreetMap (base map only).**

*Version Two*: These snap point list were created in 2012. The centre point of every road in England and Wales was taken from the Ordnance Survey Locator dataset. These points were augmented with locally relevant points of interest from the Point X dataset. Points of interest include public places, such as Parks, or commercial premises, such as Shopping Centres. Each map point was analysed to see how many postal addresses were contained in its catchment area according to the Ordnance Survey Address-Point dataset. Again, the catchments were created using Voronoi polygons. Any catchments with between 1 and 7 postal addresses were discarded to protect privacy. The remaining points were provided to police forces for human assessment; a small number of additions and deletions were made to make map points more locally relevant.

Our main source of information regarding these snap points are police.uk itself and its accompanying data website: data.police.uk (see data.police,uk n.d and its web archived version in 2013). Information on Version One was inferred from Smith and Heath (2014) and cross-referenced with information from older versions of data.police.uk. We inferred the timing of the version from looking at the dates that unique points were first recorded on police.uk (see data section). For both versions, the locations of snap-points are based on the location of static urban features. The centre of streets form the vast majority of snap-points in Version Two and the entirety of Version One.

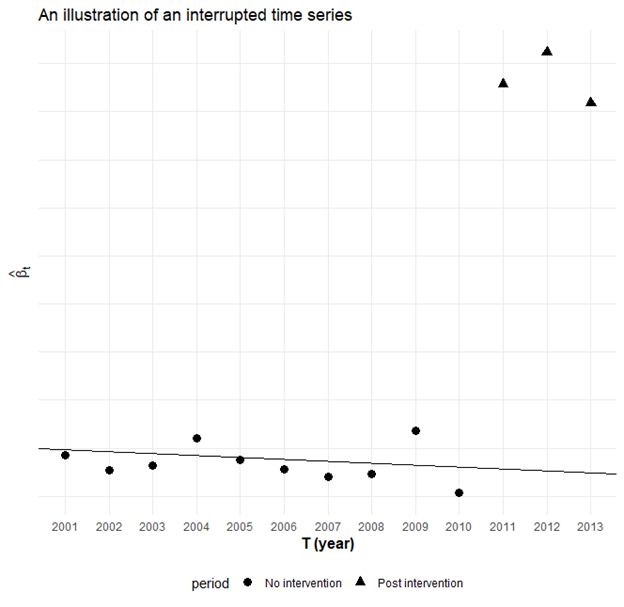
In theory, since police.uk used Ordnance Survey data, we can recreate the majority of snap-points in their master snaps list. In practice, given the density of crime in urban regions and a long enough timeframe, all the unique snap points on police.uk will eventually be used. So we can infer the master snaps list by counting the number of unique locations used by police.uk during a given period. Using this method, we can recover 96% of all the snap-points used in the 2012 and the vast majority used in Version One (see supplementary materials for more details).

The geomasking routine produces at least two features uniquely associated with police.uk: - the number of snap-points in an area - the discrepancy in crime counts as recorded by police.uk and ‘real’ police force data in the area (i.e. geomasking error)

Furthermore, since police.uk’s geomasking algorithm is documented, we can deploy it on policing data for years before the launch of police.uk. Together these pieces of information form the core of our research design.

# Research design

Members of the public or other bodies (e.g. estate agents) are not aware of the actual location of crimes as recorded by the police. Also, street-level crime maps for South Yorkshire did not exist before police.uk. Broader areal level maps have existed since 2008; notably the crimemapper website. The research design exploits the fact that key features of police.uk, namely its snap point list and geomasking errors, did not exist before 2011 and had no causal impact on house prices before that date. Any statistical association between these features and house prices are due to confounding (e.g. common causes). After the launch of police.uk in 2011, these features could have a causal effect on house prices. Assuming that the relationship between confounders and these key features are constant over time, changes in the association between these key features and house prices before and after the launch of police.uk will be indicative of causality. In essence, we are employing a form of interrupted time series analysis. Figure 2 demonstrates the intuition behind our research design.

 **Figure 2: An illustration of an interrupted time series using police.uk data**

is the associations (e.g. correlation coefficient, regression slope) between house prices and either number of snap-point or geomasking error. A change in after the launch of police.uk in 2011 indicates causality. We can also check for a constant trend in before the intervention (e.g.  doesn’t wildly fluctuate for no reason).

First, we can test for the impact of police.uk on house prices if we knew the secret list of snap points used by police.uk. These snap points are based on a subset of primarily static urban features (mainly the centre of residential roads) as recorded in 2012 (and maybe 2011). Before 2011, the number of potential snap points around a house could not casually affect its selling price. It will have an association with selling price because the snap points are constructed from Ordinance Survey data which in turn reflect urban features (e.g street centres). Both could have different effects on house prices. After the launch of police.uk, the number of potential snaps around a house could have a causal effect on house prices as areas with more snaps will have more crimes shown. For example, at the extreme end, houses in areas with no potential snap points nearby will have no crimes reported on the police.uk crime map. Police.uk could have an effect outside of the information shown on its crime maps. Most notably, it could affect the credibility of other sources of crime information. For instance, more accurate crime reporting on the website could reduce the impact of other sources of word of mouth or older crime maps. On the other hand, errors in the location of crimes on police.uk weakens the relationship between crimes and house prices. We will show the early results of an initial analysis here using data from South Yorkshire. As we mentioned before, most snap-points can be inferred from police.uk data.

Second, using-snap points alone, we cannot disentangle the impact of crime shown on crime maps from the effect of police.uk on other causal factors. We will use geomasking error around a property as a source of variation in the crime shown on a crime map to resolve this issue. If the level of geomasking error was completely random, we could use it to estimate the effects of police.uk on house prices. However, geomasking errors are not entirely random. Due to the rule governing the creation of the snap points, high levels of geomasking error are associated with proximity to urban features (e.g. high density housing, centre of roads). These features will also affect house prices.

As stated above, these urban features are primarily static over time and space, and therefore their confounding effects tend to remain invariant over time. Furthermore, the geomasking technique used by police.uk can be replicated since it is well documented and based on a list of static snap-points. The latter is secret, but we mostly infer the coordinates of these snap-points from public data. Therefore, we can mimic what public crime information would have looked like on the website if it had launched a year or two earlier. Then we can calculate geomasking errors around an area for periods before police.uk’s launch.

As with potential snaps points, we would look for differences in the association between geomasking errors and in periods before and after police.uk’s launch. Our preferred estimator would compare houses sold in areas with different levels of geomasking error but the same level of real crime. This is to block other causal pathways between police.uk and house prices that are not related to the information shown on the crime maps (see next section).

## Methodology and Statistical analysis

Begin with a limited version of the causal relationship between snap points around a house, crimes on police.uk around a house, selling price and confounders (many of which we observe). Let:

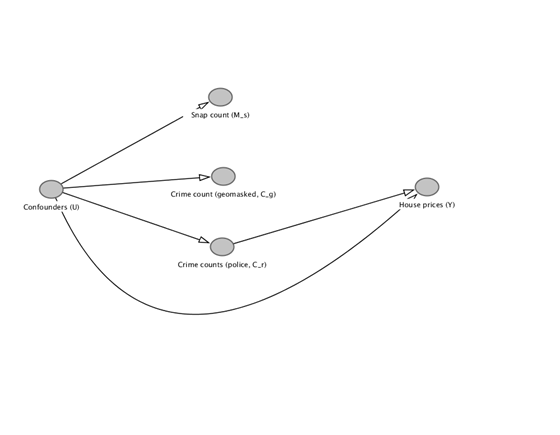
* selling price of a house (logged). can be demeaned to adjust for inflation but this is irrelevant later (e.g. due to the inclusion of an intercept in OLS models ).
* Total crime counts around a house using police.uk (i.e. geomasked crime count). For records before Dec 2010, we use South Yorkshire Police’s geocoded data, our inferred-snap list, and details from data.police.uk to create what data would have been on police.uk if it had launched earlier (see data section).
* Total crime counts around a house using police force records. Although errors can exist in the police data, we assume this is the real crime count for simplicity. We do not believe this will adversely affect our design.
* Sum of snaps around a house
* confounding factors affecting and other variables. Some are observed, and some are unobserved.
* represent entire information sets/ datasets with a subscript denoting different information.
  + is the secret list of snap-points used by police.uk. We know how was constructed (see data section).
  + represents crimes, where they occur and crime type. The level of crimes around a house (summarised by ) affects house prices in various ways (e.g. property destruction).
  + represents houses and their characteristics, including their location.

Imagine two period: one before the launch of police.uk ( = 0) and one afterwards (). For now, assume that refers to the year 2010 and is the year 2011. Furthermore imagine these counterfactuls: let be a world where police.uk’s website existed during period and be a world it didn’t.

We know for certain that key features of police.uk, namely snaps and crimes shown ( and ), has no causal impact when . If the website did not exist – its contents cannot affect house prices. When , the statistical association between and due to confounder(s) . To discover whether police.uk affected house prices, we compare the statistical association between i) and and ii) and . For example, we can check if:

However, we cannot ever observe a world in which police.uk did not exist in 2011 (i.e. ). But we do observe data from the year 2010 when police.uk did no exist (i.e. ). We can substitute data from 2010 for the data from the counterfactual .

The below Directed Acyclic Graph (DAG) represents our core assumptions about causal relationships in a world where police.uk did not exist (Pearl 2009). A more extensive version is shown in the appendix.

 **Figure 3: DAG/ Path diagram of causal relations**

We assume the existence of police.uk only affected house prices through i) the information shown on the website (modifying ) and ii) changing the relationship between information outside the website and house prices (, ). Other causal relationships are unchanged, and this forms the basis for our assumption tests later.

## Estimators (RQ1)

From the DAG, at we can see that the statistical correlation between and is simply due to confounding. If the police.uk affected house prices, then this correlation will change at . Therefore we can answer RQ1 by checking the statistical relationship between house prices and the number of potential nearby snap-points .

**Estimator 1A: Non-parametric**

Calculate the correlation (i.e. spearman’s rank) between and at and separately. Under the null hypothesis the correlation would be identical for both time periods. We can use bootstrapping or permutation tests to get standard errors.

**Estimator 1B: Parametric** Assuming that a linear functional form is close enough approximation, estimate the following model using OLS:

Where denotes parameters to be estimated from the data. The interaction term denotes a change in the statistical relationship between and over time. Under the null hypothesis that police.uk had no effect, we would expect to be zero.

**Estimator 1C: Parametric time series** Assume that a linear functional form is close enough approximation. Then redefine to be an index number for year (e.g.  if the year is 2011). For every year, estimate the following model using OLS:

The statistical association between and () may change over time. If that change was consistent (e.g. greater effect sizes due to inflation) then, for every year before 2011, we can fit the following model using OLS:

We can use the model to estimate . This is what would have been if police.uk did not launch in 2011. Under the null hypothesis:

## Estimators (RQ2)

To answer RQ2, we can directly estimate the effect of the information shown on its website by exploiting the discrepancies in , which is what the public observes, and the actual crime count . By comparing areas with the same underlying crime levels () but different crime counts as shown on police.uk, we can estimate the causal effect of on house prices.

This is only possible if either police.uk’s had no effect on house prices outside its website () or police.uk only affected the impact of other crime-related factors on house prices. For instance, if police.uk affected the credibility of our sources of crime information like word of mouth or caused the shutdown of other crime maps. In this instance, we assume that this is entirely captured by the causal relationship and that other pathways are unchanged.

Under such conditions, we can compare prices for houses sold in areas with the same underlying crime levels () but different crimes counts () as shown on police.uk. From the DAG, the observed statistical relationship between and conditional on is a function of both confounding and collider bias ( is a collider). If had no effect on house prices then this statistical relationship should remain exactly the same at and .

**Estimator 2A: Non-parametric** Check : the expected value of given , and . Since and are continuous counts, we will have to band these into categories (based on say quartiles). For example, a house is in band if its values of and fall in the lowest quartile.

For each band, calculate difference mean over time:

Under the null hypothesis, this value should be equal to zero. Given the large number of potential bands, we may have to correct for multiple tests (e.g. Bonferroni correction).

**Estimator 2B: Semi-parametric using matching** Match each case (i.e. house sold) in to cases in based on and . Discard unmatched cases. Ideally, the remaining data in and should have identical distributions of and (i.e. ).

Assuming approximate linearity, estimate the following using OLS:

Under the null hypothesis that has no effect, is equal to zero. is the estimate for the impact of a one-unit increase in on house prices. We can also replace with dummy variables for quartiles.

## Alternative estimators

We can further extend estimators 1B, 1C, and 2B using areal fixed effects. We will use lower super output area (LSOA) fixed effects. This is because public information at levels lower than LSOA is rare due to small area estimation and privacy issues. For example, the much-used index of multiple deprivation only reports deprivation at the LSOA level. We are almost certain that the older crime maps did not contain information at a lowest geographical scale than LSOA (with one potential exception for West Yorkshire). In addition, fixed effects control for time-variant confounders. As we mentioned previously, the major sources of confounding are due to static urban features and arise due to how police.uk created the snap-point list.

For example, the alternative estimator with FE for 1B would be:

Where are the LSOA fixed effects. The quantity of interest (i.e. ) and all other things remain unchanged. Standard errors have to be adjusted for fixed effects.

Instead of using the log of house prices as our outcomes , we can try to use house prices without any transformations. This is only a useful estimator if inflation between and is negligible. Else if inflation is non-negligible but consistent, then estimator 1C (i.e. time series) will remain unbiased.

Another alternate estimator uses the relationship between crimes shown on police.uk and to answer RQ1. Replace with in estimators 1A - 1C and change the relevant quantities of interest. Everything else remains the same. There are data limitations based on how many years of geocoded data (before 2011) have been archived by police forces. SYP can only guarantee that data from 2010 is available.

For estimator 1C, which is relevant in case of trends over time, we can explore more elaborate time series models with different techniques for identifying structural breaks. From our early finding (see appendix), we believe that the simple OLS time series is a good baseline.

Our statistical tests are designed to reject the null hypothesis. Failure to reject the null hypothesis does not mean that the null hypothesis is true (i.e. police.uk did not affect house prices). An alternative approach is to calculate the probability that the true effect size is higher than some substantial amount. For example, if we can check the probability that a standard deviation change in causes an increase in house prices higher than 2%. This would involve some Bayesian statistics.

## Sensitivity and Robustness tests

From the DAG in Figure 5, we can infer most of our key assumptions which are:

1. The effect of confounders on crimes shown on police.uk does not change over time
2. The effect of confounders on potential nearby snaps does not change over time
3. The distribution of confounders do not change over time
4. The distribution of selling prices did not change for other reasons. To account for rising average house prices, we can demean (or adding an intercept term in regression models).

For RQ2, we need the additional assumption:

1. Between and , any change in the effect of on is entirely mediated by the real crime count .

Many of these assumptions involve , confounder which are common causes of , and . We do observe confounders such type of housing sold, existence of nearby amenities, and so forth. For the confounders that we do observe we can check if:

by comparing univariate and multivariate statistics between and . We can also do similar tests to check .

We adopt a kitchen-sink approach to testing: get as many possible variables as possible and check for changes in correlation (or other statistics) over time. We outline some potential candidates for in the data section.

We can also do a pre-intervention test: find another period before where police.uk did not launch and check that the same assumptions are met (e.g. ). For example, can be the year 2009. Furthermore we can check that our estimators (e.g. 1A - 1C) give the expected result of no effects.

We have no statistical way to test the assumption that between and , any change in the effect of on will be entirely mediated by the real crime count . There may be other mediating pathways that do not travel along . For example. is a mediator whose effect on house prices differs between and $T = 0 $. If there is a single causal pathways between and that does not lies along the causal pathway then this would violate our assumptions.

We do not expect spatial autocorrelation to greatly affect our results. However, we will test for spatial autocorrelation (i.e. checking Moran’s I) and adjust estimates accordingly.

Finally, we can resort to investigative work to uncover evidence that may refute any of our assumptions. In particular, we will look for other ways that the police.uk crime maps may have affected house prices in non-crime related ways.

# Data

## Data sources / Datasets used

The HM Land Registry Price Paid dataset is a publicly available dataset of properties sold in England and Wales since 1995. The dataset contains information on price sold, address, house type and other features of the property sold. The dataset excludes certain types of transactions, such as inheritance and discounted transactions (e.g. discounted sales of social housing under the ‘Right to Buy’ scheme). The dataset can be accessed at: <https://www.gov.uk/government/collections/price-paid-data>. The coordinates of a property are derived from the coordinates of its postcode as recorded in the ONS National Statistics Postcode Lookup (ONS, 2021).

Archival data from police.uk are publicly available from the police.uk data site (<https://data.police.uk/data/archive/>). Other information, such as police force boundaries, are also contained on the website. We use the earliest archival extract of police.uk which contains data on crimes from Dec 2010 to Dec 2013. In general, police.uk keeps excellent documentation on archival data and changes made to its website and data manipulation. We also cross-referenced the historical police.uk website using the Wayback machine, which is an archive of websites.

For South Yorkshire Police crime data, we will use the same data source sent to the Home Office and ultimately processed by police.uk. We expect data from at least the year 2010 to be available.

We infer the snaps used by police.uk from the unique crime locations shown on police.uk during these periods covered by versions one and two (see supplementary materials for more details). In our inferred dataset, we have 734,000 snaps in version 2, roughly 96% of all the snaps in use by police.uk during this period. The inferred snaps are much lower in version one (~462k); we do not know how many snaps were used in this version. Ideally, we would like to use the entire master snap point data to mitigate against all measurement errors. However, for our estimators, it only matters that the causal relationship between our inferred snap points and confounders remains constant during the period under study.

To test the plausibility of the research design, we use a variety of data sources. In short, we want to use this information to check that the distribution of potential confounders is the same before and after the launch of police.uk. Ideally, confounders would be entirely uncorrelated with the number of snaps and errors in geomasking. We aim to use a kitchen-sink approach: assume that any variable can be a confounder and test it all. We expect to use:

* Food agency standards ratings. This is points data that contain information on business type, address, latitude and longitude alongside the most recent hygiene rating of an establishment
* Radon readings per 1km square grid. This is a public dataset.
* Aggregated data (LSOA) on households from the British Household panel
* Access to Healthy Assets and Hazards. This is data from the Consumer Data Research Centre (CDRC). These are areal (LSOA) measures of how ‘healthy’ a neighbourhood including the mean distance to: – retail environments (e.g. fast food outlets) – health services (e.g. GP offices, hospitals etc) – physical environment (e.g. green space) – air quality
* Levels of homeownership in the area (output area, lower super output area)

## Data restrictions

Police force data is not directly available for research. Researchers need to contact and negotiate access with individual police forces. Other data sources mentioned are publicly available or else require a free registered account. In the case of Ordinance Survey product (OS), researchers may need to purchase data from OS.

We inferred the master list of snap-points from public domain data from data.police.uk. The inferred list cannot be used for reverse geomasking (i.e. to reveal the exact location of crimes and identify victims). We have explained exactly how to recreate this list using public data and have share our code on Github [link redacted for peer review] and as supplementary materials for this protocol.

We will not bring any house price data (i.e. outcomes information) into South Yorkshire Police’s settings until after the protocol is finalised to protect the validity of the study protocol.

### Variables/ Codebook

The main variables used are: - police.uk variables (points data) - Date of offence/ incident (truncated to year and month only) - Home Office Offence Code - Latitude and Longitude (WGS84, to be converted to Easting and Northing OSGB36) - South Yorkshire Police force variables (to be confirmed) - Date of offence/ incident - Home Office Offence Code - Easting and Northing (OSGB36) - Other contexts (e.g. free text information about location) - Land registry price paid data - Date of transfer. Date when the sale was completed, as stated on the transfer deed. - Price paid - Postcode (joined to coordinates of postcode centroid via the ONS master postcode lookup) - Type of property (e.g. Flats, Detached housing etc) - Old/New. Whether a property was newly built or an established residential building - Freehold or Leasehold Price Paid Data (PPD) category. Relates to type of price paid data and data recording changes over time. We use category A which forms the bulk of the dataset and is available from 1995. Category B transactions are only recorded since October 2013. These include transfers under a power of sale/repossessions and buy-to-lets.

For every residential property sold, we can derive the following variables: - number of potential snap points nearby (based on a particular snap list version) - number of crimes and incidents nearby in the past three months (police force recorded). - number of crimes and incidents nearby in the past three months (police.uk recorded) - number of crimes and incidents nearby in the past three months of available police.uk data.

As mentioned early, there is a lag in police.uk data, homebuyers buying in February can only access data up to January (and maybe even less recently than that).

Data on crimes in the prior three months is chosen based on Braakman’s research (2017). Nearby is defined as within 150/300/500m; our preferred distance is 150 because the lowest level of points data on police.uk appears at a specific zoom level. At that zoom level, the scope of the interactive map on police.uk roughly covers a 300m by 300m square.

### Missing data

There are almost no missing values in our main variables. Where missing values exists, we will perform list-wise deletion (i.e. get rid of cases with missing fields).

### Sample

Our unit of analysis are residential properties that were sold in England and Wales. For RQ1, we restrict the sample to data on houses: - sold before 2014 - sold in England and Wales. We’ll focus on South Yorkshire in the first instance before extending to cover all other areas. - which are not newly built - PPD category (see variable) is category A. - not in the top or bottom 2% of price sold (trimming outliers)

For RQ2, we add further restrictions due to data limitations and look at houses: - sold in the years 2010 to 2013 (or earlier if data allows) - sold within South Yorkshire Police’s force boundaries

For the protocol, the study period is limited to years before 2014. 2014 was chosen as the limit for several reasons. First, we want to investigate in detail what happened in the years around the launch of police.uk. Given our small resources, we have to constrain ourselves to a few years to investigate. Second, we want to make sure that the housing market and policy remain relatively stable in the years just before and after police.uk. Three years after the launch of police.uk seemed to be sensible. Third, in 2013 a scheme aimed at helping first time home buyers called Help to Buy was launched. The first phase of the scheme only affected buyers looking to buy newly built homes (which are excluded from our sample). The second phase was launched in October 2013 and did not just affect newly built homes. This scheme increased demand in the housing market, but it’s not apparent how that will affect the validity of our design. As a precaution, we limit our protocol analysis to years up to 2013 (i.e to balance credibility and statistical power).

For RQ2, after restricting our sample, we have roughly 10,000 houses sold every year between 2010 and 2012 (12,000 sold in 2013). The mean house price is virtually unchanged while the price variance increased in 2013.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Year** | **N** | **Mean Price** | **Mean Ln Price** | **Sd Price** | **Sd Log Price** |
| **2010** | 9598 | 143983 | 11.7 | 94415 | 0.530 |
| **2011** | 9798 | 135934 | 11.7 | 96959 | 0.521 |
| **2012** | 9915 | 141522 | 11.7 | 94802 | 0.532 |
| **2012** | 12064 | 142829 | 11.7 | 118282 | 0.546 |

**Table 1: Summary of house price data**

To do a power analysis, we need to know the variance of the ‘treatment’ variable: the sum of potential snaps points for RQ1 (M\_S) and the crime counts on police.uk (conditional on real crime counts) for RQ2. For RQ2, we cannot know this statistic before accessing the data. For RQ1, we already have early results for South Yorkshire. From those results, we are confident in the statistical power of our estimator when applied to the entirety of England and Wales (see supplementary materials).

# Other considerations

## Study limitations and risks

A data protection plan formed between SYP and Sheffield University minimises the disclosure of personal data.

police.uk’s snaps database remains a secret to this day. This guarantees that the snap point database cannot causally affect house prices except via police.uk’s website and API. However, we have to replicate the snap point database using public information. We can infer the majority of the snap-points through police.uk data but our inferred database can still contain errors. We can test the accuracy of our inferred snap database using real police force data.

The scope of this study is only limited to a number of years, mainly 2010 - 2013. For RQ2, we can only study the effects for South Yorkshire.

## Ethics

This project has been approved by the University of Sheffield (self-declaration: 043654). It only uses existing data that has been robustly anonymised.

## Data protection

All personal data will be stored within SYP setting. No personal data will leave these settings. All other data is publicly available. Only aggregated results will be taken out of the setting and cleared by SYP beforehand.

# Bibliography

Braakmann, Nils. 2017. ‘The Link between Crime Risk and Property Prices in England and Wales: Evidence from Street-Level Data’. Urban Studies 54 (8): 1990–2007. <https://doi.org/10.1177/0042098016634611>.

Ceccato, V. and Wilhelmsson, M. 2011. ‘The impact of crime on apartment prices: Evidence from Stockholm, Sweden’. Geografiska Annaler: Series B, Human Geography, 93, 81–103.

Chainey, Spencer, and Lisa Tompson. 2008. Crime Mapping Case Studies: Practice and Research. London, UK: John Wiley & Sons, Ltd.

Chainey, S., and L. Tompson. 2012. ‘Engagement, Empowerment and Transparency: Publishing Crime Statistics Using Online Crime Mapping1’. Policing 6 (3): 228–39. <https://doi.org/10.1093/police/pas006>.

data.police.uk. n.d. ‘About | Data.Police.Uk’. Accessed 18 March 2022. <https://data.police.uk/about/>. Older web archived version are here on the Wayback machine: <https://web.archive.org/web/20130705204237/https://data.police.uk/about/>

Feng, A, and G Graetz. 2015. A Question of Degree: The Effects of Degree Class on Labour Market Outcomes. Centre for Economic Performance Discussion Paper. London: Centre for Economic Performance.

Gibbons, Steve. 2004. ‘The Costs of Urban Property Crime’. The Economic Journal 114 (499): F441–63. <https://doi.org/10.1111/j.1468-0297.2004.00254.x>.

Herbert, Nick. 2011. ‘Crime Maps Aren’t a Gimmick. They’ll Make the Police Accountable’. The Guardian, 10 February 2011, sec. Opinion. <https://www.theguardian.com/commentisfree/2011/feb/10/crime-maps-policing>.

Ihlanfeldt, Keith, and Tom Mayock. 2010. ‘Panel Data Estimates of the Effects of Different Types of Crime on Housing Prices’. Regional Science and Urban Economics 40 (2–3): 161–72. <https://doi.org/10.1016/j.regsciurbeco.2010.02.005>.

ONS (2021) Office for National Statistics, National Statistics Postcode Lookup (February 2021). Available at: <https://geoportal.statistics.gov.uk/datasets/7606baba633d4bbca3f2510ab78acf61/about> (Accessed on 18th April 2022)

Pearl, Judea. 2009. Causality: Models, Reasoning and Inference. 2nd ed. Cambridge, UK: Cambridge University Press.

Pinchbeck, Edward, Sefi Roth, Nikodem Szumilo, and Enrico Vanino. 2020. ‘The Price of Indoor Air Pollution: Evidence from Radon Maps and the Housing Market’. SSRN Electronic Journal. <https://doi.org/10.2139/ssrn.3687132>.

Quinn, Anthony, Louise Cooke, and Mark Monaghan. 2019. ‘An Exploration of the Progress of Open Crime Data: How Do Ongoing Limitations with the Police.Uk Website Restrict a Comprehensive Understanding of Recorded Crime?’ Policing and Society 29 (4): 455–70. <https://doi.org/10.1080/10439463.2017.1397149>.

Quinton, Paul. 2011. ‘The Impact of Information about Crime and Policing on Public Perceptions’, 63.

Ratcliffe, Jerry H. 2002. ‘Damned If You Don’t, Damned If You Do: Crime Mapping and Its Implications in the Real World’. Policing and Society 12 (3): 211–25. <https://doi.org/10.1080/10439460290018463>.

Sampson, F., and F. Kinnear. 2010. ‘Plotting Crimes: Too True to Be Good? The Rationale and Risks behind Crime Mapping in the UK’. Policing 4 (1): 15–27. <https://doi.org/10.1093/police/pap015>.

Semrush 2022. Traffic Analytics: www.police.uk. <https://www.semrush.com/analytics/traffic/overview/www.police.uk?searchType=domain>

Smith, Amanda M and Heath, Tom. 2014. ‘Police.Uk and Data.Police.Uk: Developing Open Crime and Justice Data for the UK.’ JeDEM - EJournal of EDemocracy and Open Government 6 (1): 87–96. <https://doi.org/10.29379/jedem.v6i1.326>.

Tompson, Lisa, Shane Johnson, Matthew Ashby, Chloe Perkins, and Phillip Edwards. 2015. ‘UK Open Source Crime Data: Accuracy and Possibilities for Research’. Cartography and Geographic Information Science 42 (2): 97–111. <https://doi.org/10.1080/15230406.2014.972456>.

Travis, Alan, and Hélène Mulholland. 2011. ‘Online Crime Maps Crash under Weight of 18 Million Hits an Hour | Police | The Guardian’. Accessed 18 March 2022. <https://www.theguardian.com/uk/2011/feb/01/online-crime-maps-power-hands-people>.