Contents lists available at ScienceDirect

Data in Brief

journal homepage: www.elsevier.com/locate/dib

Data Article

Integrated statistical indicators from Scottish linked open government data

Areti Karamanou[a,b,*], Evangelos Kalampokis[a,b], Konstantinos Tarabanis[a,b]

[a] *Center for Research & Technology HELLAS (CERTH), 6th km Charilaou-Thermi Rd., 57001 Thessaloniki, Greece*
[b] *Information Systems Lab, Department of Business Administration, University of Macedonia, Egnatia 156, Thessaloniki, 54636, Greece*

ARTICLE INFO

Article history:
Received 20 October 2022
Revised 9 November 2022
Accepted 17 November 2022
Available online 23 November 2022

Dataset link: Integrated Statistical Indicators from Scottish Linked Open Government Data (Original data)

Keywords:
Integrated statistical indicators
Linked data
Open government data
Scottish statistics
Machine learning

ABSTRACT

Open Government Data (OGD), including statistical data, such as economic, environmental and social indicators, are data published by the public sector for free reuse. These data have a huge potential when exploited using Machine Learning methods. Linked Data technologies facilitate retrieving integrated statistical indicators by defining and executing SPARQL queries. However, statistical indicators are available in different temporal and spatial granularity levels as well using different units of measurement. This data article describes the integrated statistical indicators that were retrieved from the official Scottish data portal in order to facilitate the exploitation of Machine Learning methods in OGD. Multiple SPARQL queries as well as manual search in the data portal were employed towards this end. The resulted dataset comprises the maximum number of compatible datasets, i.e., datasets with matching temporal and spatial characteristics. In particular, the data include 60 statistical indicators from seven categories such as health and social care, housing, and crime and justice. The indicators refer to the 6,976 "2011 data zones" of Scotland, while the year of reference is 2015. Data are ready to be used by the research community, students, policy

DOI of original article: 10.1016/j.bdr.2022.100355

* Corresponding author.

E-mail addresses: akarm@uom.edu.gr, a.karamanou@iti.gr (A. Karamanou), ekal@uom.edu.gr (E. Kalampokis), kat@uom.edu.gr (K. Tarabanis).

Social media: @aretikaramanou (A. Karamanou), @kalampokis (E. Kalampokis)

https://doi.org/10.1016/j.dib.2022.108779

makers, and journalists and give rise to plenty of social, business, and research scenarios that can be solved using Machine Learning technologies and methods.

Specifications Table

Subject	Decision Sciences (General)
Specific subject area	Decision and policy making using Open Government Data and Machine Learning
Type of data	Table Excel file
How the data were acquired	The data were collected from an OGD portal that disseminates official statistics using linked data technologies and a SPARQL endpoint. Multiple SPARQL queries as well as manual search in the data portal were employed to create the dataset. SPARQL is the standard query language and protocol for linked data and RDF databases. Specifically, we used SPARQL queries to find compatible datasets, i.e., datasets with matching temporal and spatial characteristics, and to retrieve the data from these datasets. Since not all datasets of the data portal are available for all years, we submitted the SPARQL query multiple times in order to find the year with the largest number of compatible variables. This was required because the aim is to use the data in Machine Learning problems that require large amounts of data. All datasets of the data portal were also manually searched to find additional compatible datasets that were referring to broader time periods. The data of the final list of compatible datasets were retrieved by submitting a second SPARQL query.
Data format	Raw Filtered
Description of data collection	Data were collected from the Scottish OGD portal* that disseminates official statistics using linked data technologies. The datasets in the portal measure statistics using dimensions (e.g., area, time), and attributes (e.g., the unit of measure). The statistical indicators of the dataset described in this article are percentages or categorical variables for (a) 6,976 “2011 data zones” of Scotland, (b) year 2015, (c) a single value for the other dimensions (e.g., “male” for the “gender” dimension).
Data source location	• Institution: The site https://statistics.gov.scot/ is managed by the Scottish Government on behalf of all producers of official statistics in Scotland. It provides statistics from a variety of organizations such as the Scottish Government, National Records of Scotland, NHS Information Services Division and Transport Scotland. • City/Town/Region: Edinburgh • Country: Scotland • Latitude and longitude (and GPS coordinates, if possible) for collected samples/data: n/a The following raw data sources were used: 1. Travel times to key services by car or public transport 2. Fire - Type of Incident 3. Child Benefit - Children 4. Child Benefit - Families 5. Children in Low Income Families 6. Land Area (based on 2011 Data Zones) 7. Urban Rural Classification (6-Fold) 8. Age of First Time Mothers 9. Ante-Natal Smoking 10. House Prices 11. Dwellings per Hectare 12. Dwellings by Type 13. Household Estimates 14. Scottish Index of Multiple Deprivation - Crime Indicators

(continued on next page)

	15. Scottish Index of Multiple Deprivation - Employment Indicators 16. Scottish Index of Multiple Deprivation - Health Indicators
Data accessibility	***Primary data*** Repository name: Scotland's official statistics Data identification number: n/a Direct URL to data: https://statistics.gov.scot/ ***Secondary data*** Repository name: Zenodo Data identification number: 10.5281/zenodo.6901000 Direct URL to data: https://doi.org/10.5281/zenodo.6901000 Direct URL to SPARQL queries used to acquire data: https://doi.org/10.5281/zenodo.7304041
Related research article	A. Karamanou, E. Kalampokis, K. Tarabanis (2022). Linked Open Government Data to Predict and Explain House Prices: The Case of Scottish Statistics Portal, Big Data Research, Volume 30, 100355, ISSN 2214-5796, https://doi.org/10.1016/j.bdr.2022.100355.

* https://statistics.gov.scot.

Value of the Data

- Policy makers, government agencies, and researchers can exploit the integrated statistical indicators to explore the potential of OGD using Machine Learning methods (e.g., [1]).
- Integrated statistical indicators and their descriptive analytics can be further analyzed to identify new patterns, trends, and data anomalies.
- Integrated statistical indicators can be employed to define and address problems related to the health of citizens, economics, environment, social problems, and others using Machine Learning.
- Real estate agencies can use the integrated statistical indicators to enhance their Machine Learning models that predict mean house prices.
- Integrated statistical indicators can be also combined with other data that describe, for example, structural features of houses and in various formats, such as satellite images in order to estimate the prices of individual houses (e.g., [2,3]).
- Researchers can use the integrated statistical indicators to create and benchmark Machine Learning models against results already known or benchmark the performance of different types of Machine Learning algorithms for solving the same type of problem (classification, regression, etc.)

1. Objective

Open Government Data (OGD), including statistical data, such as economic, environmental and social indicators, are data published by the public sector for free reuse. These data have a huge potential when exploited using Machine Learning methods. Linked Data technologies facilitate retrieving integrated statistical indicators by defining and executing SPARQL queries. However, statistical indicators are available in different temporal and spatial granularity levels as well using different units of measurement. The objective of this data article is to describe the integrated statistical indicators that were retrieved from the official Scottish data portal in order to facilitate the exploitation of Machine Learning methods in OGD. Multiple SPARQL queries as well as manual search in the data portal were employed towards this end. The resulted dataset comprises the maximum number of compatible datasets, i.e., datasets with matching temporal and spatial characteristics. The dataset can be used by the research community, students, policy makers, and journalists and give rise to plenty of social, business, and research scenarios that can be solved using Machine Learning technologies and methods.

2. Data Description

Open Government Data (OGD) are data published by the public sector in open and reusable formats without restriction or charge for their use by society [4]. OGD have a huge potential especially when combined with AI technologies to bring new and fruitful insights [5]. OGD is a political priority the last decade in many countries in order to harness multifaceted benefits including enhancing evidence-based policy making and stimulating economic growth. As a result, a large number of public authorities and National Statistical Institutes internationally have already been publishing their data in their official data portals (e.g., https://www.data.gov in the U.S. and www.dati.gov.it in Italy). However, the current scarce use of OGD has impeded reaching the full potential of OGD [6].

A large part of OGD is of statistical nature, meaning that it consists of highly structured numeric data and multidimensional, i.e., a measure is described based on multiple of dimensions. They also usually concern aggregated data that monitor various indicators such as economic indicators (e.g., labour productivity), environmental indicators (e.g., greenhouse gas emissions), and social indicators (e.g., employment deprivation) across countries.

OGD are often offered by various official data portals (e.g., Scottish OGD portal that disseminates official statistics using linked data technologies[1]) in a standardized manner as Linked Open Government Data (LOGD), hence facilitating their reusability and interlinking [7]. As a result, they significantly contribute towards creating powerful predictive models. However, interoperability conflicts have not been yet addressed and hamper combining and analysing LOGD [7,8].

Integrated data described in this article are LOGD coming from the Scottish OGD portal that disseminates official statistics using linked data technologies. The portal currently hosts 314 datasets covering various societal and business aspects of Scotland classified into 18 themes. Datasets are provided at different levels of spatial granularity in Scotland starting from the level of postcodes to the level of council areas. Users can navigate through the data portal to view and retrieve data as tables, maps, and charts or download them in various formats (e.g., html, json, csv), or, alternatively, retrieve them as linked data by submitting flexible queries to the SPARQL endpoint[2] released by the portal.

Integrated data are included in a single file, namely statistical_indicators.csv. The file contains 6,976 rows and 61 columns, i.e., 425536 data points. Each raw is a data zone, while the columns include the 60 statistical indicators coming from 16 datasets plus 1 column with the name of the data zone. There are 6037 null values in data (1.4%). Table 1 presents and describes the variables and their type. The majority of the variables are numeric. One (CIF) is integer and one (Urban Rural Classification) is categorical. All data are licensed under the Open Government License (OGL) version 3.0[3]. OGL v3.0 allows users to freely copy, publish, distribute, transmit, and adapt the data.

In addition, Tables 2 and 3 present the descriptive statistics for the categorical, and integer and numeric variables respectively.

[1] https://statistics.gov.scot

[2] https://statistics.gov.scot/sparql

[3] https://www.nationalarchives.gov.uk/doc/open-government-licence/version/3/

Table 1
Variables description.

a/a	Statistical indicator	Type	Description
1	Travel times (minutes) to GP surgeries by public transport	Numeric	Drive times in minutes to GP surgeries by public transport at data zone level. Average drive times are generated by generating drive times for each Census Output Area (designed specifically for statistical purposes) and then calculating a population weighted average for each data zone.
2	Travel times (minutes) to post office by public transport	Numeric	Drive times in minutes to post offices by public transport at data zone level. Average drive times are generated by generating drive times for each Census Output Area (designed specifically for statistical purposes) and then calculating a population weighted average for each data zone.
3	Travel times (minutes) to retail centre by public transport	Numeric	Drive times in minutes to retail centres by public transport at data zone level. Average drive times are generated by generating drive times for each Census Output Area (designed specifically for statistical purposes) and then calculating a population weighted average for each data zone.
4	Travel times (minutes) to petrol station by car	Numeric	Drive times in minutes to petrol stations by car at data zone level. Average drive times are generated by generating drive times for each Census Output Area (designed specifically for statistical purposes) and then calculating a population weighted average for each data zone.
5	Travel times (minutes) to post office by car	Numeric	Drive times in minutes to post offices by car at data zone level. Average drive times are generated by generating drive times for each Census Output Area (designed specifically for statistical purposes) and then calculating a population weighted average for each data zone.
6	Travel times (minutes) to GP surgeries by car	Numeric	Drive times in minutes to GP surgeries by car at data zone level. Average drive times are generated by generating drive times for each Census Output Area (designed specifically for statistical purposes) and then calculating a population weighted average for each data zone.
7	Travel times (minutes) to primary school by car	Numeric	Drive times in minutes to primary schools by car at data zone level. Average drive times are generated by generating drive times for each Census Output Area (designed specifically for statistical purposes) and then calculating a population weighted average for each data zone.
8	Travel times (minutes) to secondary school by car	Numeric	Drive times in minutes to secondary schools by car at data zone level. Average drive times are generated by generating drive times for each Census Output Area (designed specifically for statistical purposes) and then calculating a population weighted average for each data zone.
9	Travel times (minutes) to retail centre by car	Numeric	Drive times in minutes to retail centre by car at data zone level. Average drive times are generated by generating drive times for each Census Output Area (designed specifically for statistical purposes) and then calculating a population weighted average for each data zone.
10	Chimney fires (ratio)	Numeric	The rate of fires in chimneys in data zone level. Data is provided for the total number of fires, including accidental and deliberate causes. Data is entered by the Scottish Fire and Rescue Service into the Incident Recording System.
11	Dwelling fires (ratio)	Numeric	The rate of fires in dwelling in data zone level. Data is provided for the total number of fires, including accidental and deliberate causes. Data is entered by the Scottish Fire and Rescue Service into the Incident Recording System.
12	Other building fires (ratio)	Numeric	The rate of fires in other building fires in data zone level. Data is provided for the total number of fires, including accidental and deliberate causes. Data is entered by the Scottish Fire and Rescue Service into the Incident Recording System.
13	Other primary fires (ratio)	Numeric	The rate of other primary fires in data zone level. Data is provided for the total number of fires, including accidental and deliberate causes. Data is entered by the Scottish Fire and Rescue Service into the Incident Recording System.
14	Outdoor fires (ratio)	Numeric	The rate of outdoor fires in data zone level. Data is provided for the total number of fires, including accidental and deliberate causes. Data is entered by the Scottish Fire and Rescue Service into the Incident Recording System.
15	Refuse fires (ratio)	Numeric	The rate of refuse fires in data zone level. Data is provided for the total number of fires, including accidental and deliberate causes. Data is entered by the Scottish Fire and Rescue Service into the Incident Recording System.

(*continued on next page*)

Table 1 (*continued*)

a/a	Statistical indicator	Type	Description
16	Vehicle fires (ratio)	Numeric	The rate of vehicle fires in data zone level. Data is provided for the total number of fires, including accidental and deliberate causes. Data is entered by the Scottish Fire and Rescue Service into the Incident Recording System.
17	Accidental chimney fires (ratio)	Numeric	The rate of accidental fires caused in chimneys in data zone level. Data is entered by the Scottish Fire and Rescue Service into the Incident Recording System.
18	Accidental dwelling fires (ratio)	Numeric	The rate of accidental fires caused in dwellings in data zone level. Data is entered by the Scottish Fire and Rescue Service into the Incident Recording System.
19	Accidental other building fires (ratio)	Numeric	The rate of accidental fires caused in other buildings in data zone level. Data is entered by the Scottish Fire and Rescue Service into the Incident Recording System.
20	Accidental other primary fires (ratio)	Numeric	The rate of other, accidental, primary fires caused in chimneys in data zone level. Data is entered by the Scottish Fire and Rescue Service into the Incident Recording System.
21	Accidental outdoor fires (ratio)	Numeric	The rate of accidental outdoor fires in data zone level. Data is entered by the Scottish Fire and Rescue Service into the Incident Recording System.
22	Accidental refuse fires (ratio)	Numeric	The rate of accidental refuse fires in data zone level. Data is entered by the Scottish Fire and Rescue Service into the Incident Recording System.
23	Accidental vehicle fires (ratio)	Numeric	The rate of accidental vehicle fires in data zone level. Data is entered by the Scottish Fire and Rescue Service into the Incident Recording System.
24	Not accidental chimney fires (ratio)	Numeric	The rate of deliberate fires in chimneys in data zone level. Data is entered by the Scottish Fire and Rescue Service into the Incident Recording System.
25	Not accidental dwelling fires (ratio)	Numeric	The rate of deliberate fires caused in dwellings in data zone level. Data is entered by the Scottish Fire and Rescue Service into the Incident Recording System.
26	Not accidental other building fires (ratio)	Numeric	The rate of deliberate fires caused in other buildings in data zone level. Data is entered by the Scottish Fire and Rescue Service into the Incident Recording System.
27	Not accidental other primary fires (ratio)	Numeric	The rate of other, deliberate, primary fires caused in chimneys in data zone level. Data is entered by the Scottish Fire and Rescue Service into the Incident Recording System.
28	Not accidental outdoor fires (ratio)	Numeric	The rate of deliberate outdoor fires in data zone level. Data is entered by the Scottish Fire and Rescue Service into the Incident Recording System.
29	Not accidental refuse fires (ratio)	Numeric	The rate of deliberate refuse fires in data zone level. Data is entered by the Scottish Fire and Rescue Service into the Incident Recording System.
30	Not accidental vehicle fires (ratio)	Numeric	The rate of deliberate vehicle fires in data zone level. Data is entered by the Scottish Fire and Rescue Service into the Incident Recording System.
31	Crime indicators (ratio)	Numeric	The percent of Scottish Index of Multiple Deprivation (SIMD) crimes per 10,000 people in each data zone. The Crime domain is based on five indicators of broad crime types: crimes of violence, sexual offences, crimes of dishonesty, vandalism, drugs offences. The domain is calculated by summing the above types of crime that are recorded within a data zone. These crimes are collectively referred to as 'SIMD crime'. SIMD crime does not include all recorded crimes. Certain crimes have been excluded because of the data quality issues, or because they are less meaningful in terms of deprivation at a neighborhood level. Some crime types are not included in the SIMD on grounds that they are targeted at businesses or are concentrated in centers of retail activity rather than in residential neighborhoods. Fraud and speeding offences are excluded because they are harder to locate geographically.
32	Children 0-15 living in low income families (ratio)	Numeric	The percentage of children 0-15 years old that live in families in receipt of Child Tax Credit (CTC) whose reported income is less than 60 per cent of the median income or in receipt of Income Support (IS) or Income-Based Jobseekers Allowance (JSA). Data zone level.
33	Children 0-19 living in low income families (ratio)	Numeric	The percentage of children 0-19 years old that live in families in receipt of Child Tax Credit (CTC) whose reported income is less than 60 per cent of the median income or in receipt of Income Support (IS) or Income-Based Jobseekers Allowance (JSA).

(*continued on next page*)

Table 1 (*continued*)

a/a	Statistical indicator	Type	Description
34	Age of first time mothers 19 years and under (ratio)	Numeric	The percentage of first time mothers who are aged 19 and under in each data zone. The 3 year aggregate shown is for financial year ending 31 March and refers to the year of discharge from hospital.
35	Age of first time mothers 35 years and older (ratio)	Numeric	The percentage of first time mothers who are aged 35 and over in each data zone. The 3 year aggregate shown is for financial year ending 31 March and refers to the year of discharge from hospital.
36	Mothers currently smoking (ratio)	Numeric	Percentage of mothers reporting smoking at ante-natal booking visit in the community or at hospital. Data are in data zone level.
37	Mothers former smokers (ratio)	Numeric	Percentage of mothers reporting they are former smokers at ante-natal booking visit in the community or at hospital. Data are in data zone level.
38	Mothers never smoked (ratio)	Numeric	Percentage of mothers reporting they have never smoked at ante-natal booking visit in the community or at hospital. Data are in data zone level.
39	Mothers not known if they smoke (ratio)	Numeric	Percentage of mothers that didn't report their smoking status at ante-natal booking visit in the community or at hospital. Data are in data zone level.
40	Low birth-weight (less than 2500g) babies (single births) (ratio)	Numeric	The percent of low birthweight (less than 2500g) babies (single births). Data are data zone level. The 3 year aggregate shown is for financial year ending 31st March and refers to the year of discharge from hospital. Low birthweight may have been resulted from being born too soon (i.e. a preterm birth), from poor intrauterine growth or from a combination of the two.
41	Employment deprivation (ratio)	Numeric	The percentage of people who are employment deprived.
42	School attendance (ratio)	Numeric	Attendance rate of pupils attending publicly funded schools. Information is taken from attendance returns linked to the September Scottish Pupil Census, for academic year 2014/2015, for publicly funded schools. The data does not include: pupils attending special schools; pupils attending independent schools; pupils educated out with the school education system (for example at home); adults attending publicly funded secondary schools.
43	Educational attainment of school leavers (score)	Numeric	The score is based on school leavers' highest level of qualification, averaged across all leavers within a data zone. The score is calculated by multiplying the highest qualification level achieved by each pupil by a corresponding factor. Level 3 qualifications are multiplied by three, Level 4 by four, Level 5 by five and Level 6 by six. For example, one pupil who leaves school with four Level 3 qualifications will score three, whilst a pupil leaving school with one Level 5 qualification will score five. The total score is then divided by the total number of school leavers in each data zone. A pupil is assigned to a geographical area based on their home address. If their postcode is missing their school's postcode is used. Data is based on an average of three years and includes all school leavers in secondary schools and special schools. Data from independent schools is not included.
44	Land area (in hectares)	Numeric	The land area in hectares based on aggregating 2011 data zone data. Land area is provided for each data zone.
45	Urban Rural Classification	Categorical	The classification is based upon two main criteria: (i) population as defined by the National Records of Scotland, and (ii) accessibility based on drive time analysis to differentiate between accessible and remote areas in Scotland. The classification includes the following categories: (1) - Large Urban Areas Settlements of over 125,000 people; (2) - Other Urban Areas Settlements of 10,000 to 125,000 people; (3) - Accessible Small Towns Settlements of between 3,000 and 10,000 people and within 30 minutes' drive of a settlement of 10,000 or more; (4) - Remote Small Towns Settlements of between 3,000 and 10,000 people and with a drive time of over 30 minutes to a settlement of 10,000 or more; (5) - Accessible Rural Settlements of less than 3,000 people and within 30 minutes' drive of a settlement of 10,000 or more; (6) - Remote Rural Settlements of less than 3,000 people and with a drive time of over 30 minutes to a settlement of 10,000 or more.

(*continued on next page*)

Table 1 (*continued*)

a/a	Statistical indicator	Type	Description
46	Comparative Illness Factor	Integer	Comparative Illness Factor (CIF) is an indicator of health conditions. Greater CIF values indicate poorer health conditions. The Scotland average CIF is 100, hence, data zones with values of CIF greater than 100 indicate poorer health conditions related to Scotland (and vice-versa).
47	Dwellings per hectare (ratio)	Numeric	The percent of dwellings per hectare in each data zone. A ’dwelling’ refers to accommodation such as a house or a flat, and includes second homes that are not let out commercially. Caravans count as dwellings if they are someone’s main home. Data on number of dwellings is obtained from the Assessors’ Portal in December each year, or the following January. The Assessors are, amongst other things, responsible for valuing each dwelling in Scotland for the purposes of assigning it to a Council Tax Band. Dwellings per hectare is calculated by dividing the total number of dwellings by the area in hectares.
48	Detached dwellings (ratio)	Numeric	The percent of detached dwellings in each data zone. A ’dwelling’ refers to the accommodation itself, for example a house or a flat, and includes second homes that are not let out commercially. Caravans count as dwellings if they are someone’s main home. Type of dwelling is based on ’attachment’, i.e., the type of property in relation to its degree of attachment to surrounding properties. This information has been aggregated into five categories: Detached; Semi-detached; Terraced; Flat, maisonette or apartment; and Not known. Data on type of dwelling is obtained from the Assessors’ Portal in December each year, or the following January. The Assessors are, amongst other things, responsible for valuing each dwelling in Scotland for the purposes of assigning it to a Council Tax Band.
49	Flats (ratio)	Numeric	The percent of flats in each data zone. Type of dwelling is based on ’attachment’, i.e., the type of property in relation to its degree of attachment to surrounding properties. This information has been aggregated into five categories: Detached; Semi-detached; Terraced; Flat, maisonette or apartment; and Not known. Data on type of dwelling is obtained from the Assessors’ Portal in December each year, or the following January. The Assessors are, amongst other things, responsible for valuing each dwelling in Scotland for the purposes of assigning it to a Council Tax Band.
50	Semi-detached dwellings (ratio)	Numeric	The percent of semi-detached dwellings in each data zone. Type of dwelling is based on ’attachment’, i.e., the type of property in relation to its degree of attachment to surrounding properties. This information has been aggregated into five categories: Detached; Semi-detached; Terraced; Flat, maisonette or apartment; and Not known. Data on type of dwelling is obtained from the Assessors’ Portal in December each year, or the following January. The Assessors are, amongst other things, responsible for valuing each dwelling in Scotland for the purposes of assigning it to a Council Tax Band.
51	Terraced dwellings (ratio)	Numeric	The percent of terraced dwellings in each data zone. Type of dwelling is based on ’attachment’, i.e., the type of property in relation to its degree of attachment to surrounding properties. This information has been aggregated into five categories: Detached; Semi-detached; Terraced; Flat, maisonette or apartment; and Not known. Data on type of dwelling is obtained from the Assessors’ Portal in December each year, or the following January. The Assessors are, amongst other things, responsible for valuing each dwelling in Scotland for the purposes of assigning it to a Council Tax Band.
52	Dwellings of unknown type (ratio)	Numeric	The percent of dwelling in each data zone whose type is not known. Type of dwelling is based on ’attachment’, i.e., the type of property in relation to its degree of attachment to surrounding properties. This information has been aggregated into five categories: Detached; Semi-detached; Terraced; Flat, maisonette or apartment; and Not known. Data on type of dwelling is obtained from the Assessors’ Portal in December each year, or the following January. The Assessors are, amongst other things, responsible for valuing each dwelling in Scotland for the purposes of assigning it to a Council Tax Band.

(*continued on next page*)

Table 1 (*continued*)

a/a	Statistical indicator	Type	Description
53	Long-term empty households (ratio)	Numeric	The percentage of dwellings in each data zone that are long-term empty properties. Long-term empty properties include properties subject to discounts and levies due to long-term empty status. The estimates are based on two Council Tax data collections carried out each year in September. Council Tax data contains information on the various discounts/exemptions awarded to each dwelling in Scotland. From these we can determine which dwellings are occupied and which are vacant or second homes.
54	Occupied households (ratio)	Numeric	The percentage of occupied households in each data zone. Occupied household is any dwelling apart from those which are vacant or second homes. The number of occupied dwellings is a good estimate of the number of households. The estimates are based on two Council Tax data collections carried out each year in September. Council Tax data contains information on the various discounts/exemptions awarded to each dwelling in Scotland. From these we can determine which dwellings are occupied and which are vacant or second homes.
55	Second-home households (ratio)	Numeric	The percentage of dwellings in each data zone that are used as second homes. Second homes are dwellings that are not someone's main residence and that are occupied for at least 25 days a year. These include self-catering holiday accommodations available to let for a total of less than 140 days per year. Second homes which are let out for 140 days or more are not included in these figures as they are classed as business so pay non-domestic rates rather than Council Tax. Each council has discretion to apply a discount of between 10% and 50% on second homes, or may choose to apply no discount. The estimates are based on two Council Tax data collections carried out each year in September. Council Tax data contains information on the various discounts/exemptions awarded to each dwelling in Scotland. From these we can determine which dwellings are occupied and which are vacant or second homes.
56	Vacant households (ratio)	Numeric	The percentage of dwellings in each data zone that are vacant households. Vacant households are dwellings that are exempt from Council Tax and are unoccupied; and dwellings which are recorded on Council Tax systems as being long-term empty properties. The estimates are based on two Council Tax data collections carried out each year in September. Council Tax data contains information on the various discounts/exemptions awarded to each dwelling in Scotland. From these we can determine which dwellings are occupied and which are vacant or second homes.
57	Households with occupied exemptions (ratio)	Numeric	The percentage of households in each data zone with occupied exemptions. Occupied exemptions are dwellings exempt from Council Tax, which are occupied. This includes: dwellings only occupied by students, armed forces accommodation owned by the Secretary of State for Defense, dwellings which are the sole residence only of people aged under 18 or people who are classed as severely mentally impaired, trial flats used by registered housing associations, and prisons. The estimates are based on two Council Tax data collections carried out each year in September. Council Tax data contains information on the various discounts/exemptions awarded to each dwelling in Scotland. From these we can determine which dwellings are occupied and which are vacant or second homes.
58	Households with unoccupied exemptions (ratio)	Numeric	The percentage of households in each data zone with unoccupied exemptions. Unoccupied exemptions are dwellings exempt from Council Tax, which are unoccupied, such as new dwellings, those undergoing repair or awaiting demolition and dwellings where the previous owner has died. The estimates are based on two Council Tax data collections carried out each year in September. Council Tax data contains information on the various discounts/exemptions awarded to each dwelling in Scotland. From these we can determine which dwellings are occupied and which are vacant or second homes.

(*continued on next page*)

Table 1 (*continued*)

a/a	Statistical indicator	Type	Description
59	Households with single adult discounts (ratio)	Numeric	The percentage of households in each data zone with single adult discounts. Households with single adult discounts are dwellings subject to a Council Tax discount of 25 per cent. This may include, for example, dwellings with a single adult, dwellings with one adult living with one or more children, or with one or more adults who are 'disregarded' for Council Tax purposes. The estimates are based on two Council Tax data collections carried out each year in September. Council Tax data contains information on the various discounts/exemptions awarded to each dwelling in Scotland. From these we can determine which dwellings are occupied and which are vacant or second homes.
60	House prices	Numeric	The mean price of houses in each data zone based on sales that are recorded/registered in the given time period whether they are cash purchases or funded by mortgages. Data zones with less than 5 sales have the number and value of sales suppressed to help minimize the risk of data disclosure and to ensure that any averages presented are based on at least 5 records.

Table 2
Descriptive statistics for categorical variable.

Statistical indicator	Unique values	Counts and percentages per value
Urban Rural Classification	6	1: 1958 (28%), 2: 2210 (31.7%), 3: 536 (7.7%), 4: 225 (3.2%), 5: 705 (10.1%), 6: 380 (5.5%), null: 962 (13.8%)

Table 3
Descriptive statistics for the integer and numeric variables.

Statistical indicator	Null values	Mean	Std	Min	25%	50%	75%	max
Travel times (minutes) to GP surgeries by public transport	962 (13.79%)	10.33	6.32	1.6	6.3	8.8	12.4	108.8
Travel times (minutes) to post office by public transport	962 (13.79%)	8.63	4.54	2	5.6	7.5	10.3	40.3
Travel times (minutes) to retail centre by public transport	962 (13.79%)	13.56	10.43	1.9	7.9	11.2	16.4	190
Travel times (minutes) to GP surgeries by car	962 (13.79%)	3.42	2.86	0.6	1.9	2.7	4	87.8
Travel times (minutes) to petrol station by car	962 (13.79%)	3.7	3.02	0.6	2.1	2.9	4.1	63.5
Travel times (minutes) to post office by car	962 (13.79%)	2.76	1.64	0.6	1.7	2.4	3.3	17.3
Travel times (minutes) to primary school by car	962 (13.79%)	2.54	3.41	0.7	1.6	2.2	2.9	186.1
Travel times (minutes) to secondary school by car	962 (13.79%)	6.17	5.2	1.2	3.7	4.8	6.7	116.1
Travel times (minutes) to retail centre by car	962 (13.79%)	5.26	6.2	0.7	2.8	4	6	190
Chimney fires (ratio)	962 (13.79%)	19.39	73.37	0	0	0	0	1183.43
Dwelling fires (ratio)	962 (13.79%)	96.45	142.95	0	0	0	146.2	1535.84
Other building fires (ratio)	962 (13.79%)	42.87	114.18	0	0	0	0	1884.25
Other primary fires (ratio)	962 (13.79%)	15.03	54.09	0	0	0	0	1257.86
Outdoor fires (ratio)	962 (13.79%)	105.11	241.11	0	0	0	133.29	4310.34
Refuse fires (ratio)	962 (13.79%)	112.63	261.49	0	0	0	138.7	6729.48
Vehicle fires (ratio)	962 (13.79%)	33.96	85.78	0	0	0	0	1715.69
Accidental chimney fires (ratio)	962 (13.79%)	19.35	73.3	0	0	0	0	1183.43

(*continued on next page*)

Table 3 (*continued*)

Statistical indicator	Null values	Mean	Std	Min	25%	50%	75%	max
Accidental dwelling fires (ratio)	962 (13.79%)	87.18	130.49	0	0	0	139.08	1094.89
Accidental other building fires (ratio)	962 (13.79%)	32.99	100.12	0	0	0	0	1749.66
Accidental other primary fires (ratio)	962 (13.79%)	19.35	73.3	0	0	0	0	1183.43
Accidental outdoor fires (ratio)	962 (13.79%)	20.78	68.26	0	0	0	0	1130.52
Accidental refuse fires (ratio)	962 (13.79%)	20.34	101.48	0	0	0	0	4771.37
Accidental vehicle fires (ratio)	962 (13.79%)	21.84	63.58	0	0	0	0	724.64
Not accidental chimney fires (ratio)	962 (13.79%)	0.04	2.2	0	0	0	0	142.05
Not accidental dwelling fires (ratio)	962 (13.79%)	9.27	39.44	0	0	0	0	540.54
Not accidental other building fires (ratio)	962 (13.79%)	9.87	44.88	0	0	0	0	1029.3
Not accidental other primary fires (ratio)	962 (13.79%)	0.04	2.2	0	0	0	0	142.05
Not accidental outdoor fires (ratio)	962 (13.79%)	84.33	218.55	0	0	0	111.73	4310.34
Not accidental refuse fires (ratio)	962 (13.79%)	92.29	219.13	0	0	0	121.65	4441.45
Not accidental vehicle fires (ratio)	962 (13.79%)	12.13	52.65	0	0	0	0	1715.69
Children 0-15 living in low income families (ratio)	1179 (16.9%)	14.12	11.05	0	5.1	11.3	20.8	63.4
Children 0-19 living in low income families (ratio)	1179 (16.9%)	14.02	10.81	0.3	5.1	11.3	20.5	63.4
Educational attainment of school leavers (score)	1126 (16.14%)	5.57	0.5	3.2	5.25	5.62	5.93	6.8
School attendance (ratio)	971 (13.9%)	94.02	1.88	81.2	92.8	94.3	95.4	98.7
Land area (in hectares)	962 (13.8%)	1194.09	5958.34	1.56	14.08	22.35	44.94	116251.04
Age of first time mothers 19 years and under (ratio)	981 (14%)	0.69	1.09	0	0	0	1	10
Age of first time mothers 35 years and older (ratio)	981 (14%)	1.51	1.61	0	0	1	2	16
Mothers currently smoking (ratio)	963 (13.8%)	14.12	11.98	0	4.55	12	22	66.67
Mothers former smokers (ratio)	963 (13.8%)	12.84	9.24	0	6.25	11.54	18.18	66.67
Mothers never smoked (ratio)	963 (13.8%)	70.1	15.73	0	59.26	70.83	82.14	100
Mothers not known if they smoke (ratio)	963 (13.8%)	2.94	5.33	0	0	0	4.55	50
Low birthweight (less than 2500g) babies (single births) (ratio)	964 (13.81%)	0.39	0.67	0	0	0	1	5
Dwellings per hectare (ratio)	962 (13.8%)	19.66	20.65	0	8.19	16.07	24.47	219.67
Detached dwellings (ratio)	962 (13.8%)	24.75	26.17	0	1.5	14.8	43.1	100
Flats (ratio)	962 (13.8%)	32.83	31.98	0	5.2	22.3	53.98	100
Semi-detached dwellings (ratio)	962 (13.8%)	21.48	16.95	0	8.2	18.9	31.2	98.8
Terraced dwellings (ratio)	962 (13.8%)	20.38	20.82	0	4.5	13.7	29.7	100
Dwellings of unknown type (ratio)	962 (13.8%)	0.54	2.17	0	0	0	0.2	58.8
Long-term empty households (ratio)	1354 (19.4%)	5.55	7.22	0	1	3	7	114

(*continued on next page*)

Table 3 (*continued*)

Statistical indicator	Null values	Mean	Std	Min	25%	50%	75%	max
Occupied households (ratio)	1354 (19.4%)	363.88	102.98	118	295	352	418	2104
Second-home households (ratio)	1354 (19.4%)	4.1	12.54	0	0	1	3	294
Vacant households (ratio)	1354 (19.4%)	11.41	10.94	0	5	9	14	132
Households with occupied exemptions (ratio)	1354 (19.4%)	11.47	38.82	0	2	4	9	1562
Households with unoccupied exemptions (ratio)	1354 (19.4%)	5.87	6.16	0	3	5	8	131
Households with single adult discounts (ratio)	1354 (19.4%)	140.92	68.65	11	93	128	176	520
Crime indicators (ratio)	1354 (19.4%)	309.03	455.52	16	112	204	366.75	14580
Employment deprivation (ratio)	1354 (19.4%)	10.06	7.09	0	5	8	14	52
House Prices	962 (13.8%)	163478.87	91903.48	20604	101720.5	142158.5	199864.5	1244910

3. Experimental Design, Materials and Methods

The integrated data described in this article are Linked Open Government Data (LOGD) retrieved from the Scottish OGD portal that disseminates official statistics using linked data technologies. We hence, used SPARQL queries to collect them. Specifically, we used two SPARQL queries; the first one to find compatible datasets and the second one to retrieve the data from the compatible datasets and create the statistical indicators.

The methodology employed for the creation of the dataset is presented in Fig. 1. We first gradually constructed a SPARQL query to find compatible datasets in the Scottish data portal (steps 1 – 4 in the Figure). Compatible datasets are those that their (a) year of reference, and (b) granularity level of the geography dimension match. We started by querying the 314 datasets available in https://statistics.gov.scot and get only the datasets for 2011 data zones (step 1). We resulted in 41 datasets. We then updated the SPARQL query to keep only the datasets that use ratio, percent, or score (step 2). We resulted in 30 datasets. We then further updated the previous SPARQL query in order to select the year. Since not all datasets of the data portal are available for all years, we submitted this SPARQL query multiple times that used various years each time in order to find the year with the largest number of compatible variables (step 3). This was required because the aim was to use data in Machine Learning problems that require large amounts of data. We selected 2015 as the year with the most compatible datasets (11 datasets). In addition, since some datasets in the data portal use a two year or three-year time period instead of one year, the last SPARQL query could not find them. We, hence, manually searched through all datasets of the portal to find them (step 4). We found 5 more datasets. For these datasets we selected the value of the time dimension that is closer to 2015. For example, in cases of datasets that refer to two years (e.g., 2014-2015, 2015-2016, etc.), we selected 2014-

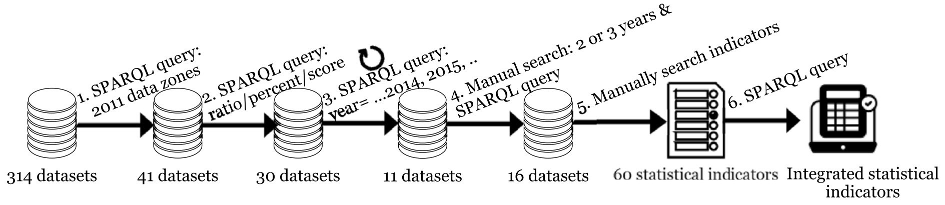
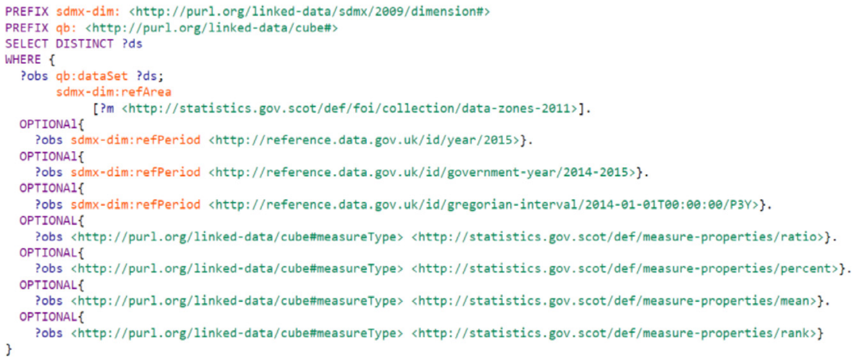
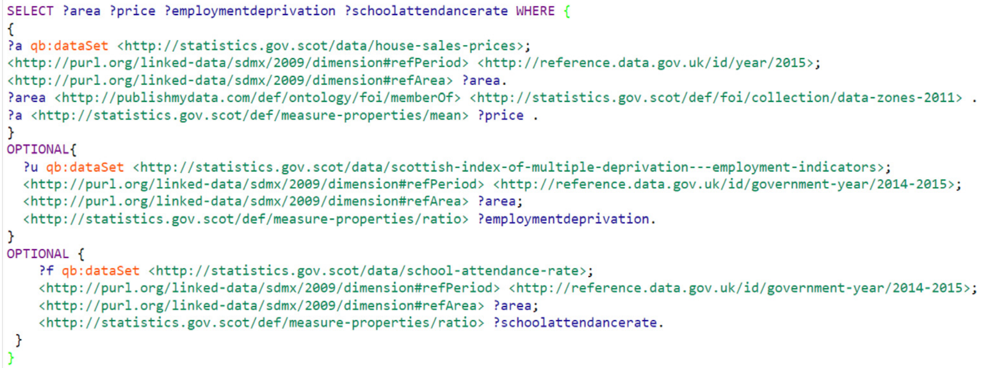

1. SPARQL query: 2011 data zones
2. SPARQL query: ratio/percent/score
3. SPARQL query: year= ...2014, 2015, ..
4. Manual search: 2 or 3 years & SPARQL query
5. Manually search indicators
6. SPARQL query
314 datasets
41 datasets
30 datasets
11 datasets
16 datasets
60 statistical indicators
Integrated statistical indicators

Fig. 1. The methodology of the creation of the dataset.



```
PREFIX sdmx-dim: <http://purl.org/linked-data/sdmx/2009/dimension#>
PREFIX qb: <http://purl.org/linked-data/cube#>
SELECT DISTINCT ?ds
WHERE {
  ?obs qb:dataSet ?ds;
       sdmx-dim:refArea
            [?m <http://statistics.gov.scot/def/foi/collection/data-zones-2011>].
  OPTIONAl{
    ?obs sdmx-dim:refPeriod <http://reference.data.gov.uk/id/year/2015>}.
  OPTIONAl{
    ?obs sdmx-dim:refPeriod <http://reference.data.gov.uk/id/government-year/2014-2015>}.
  OPTIONAl{
    ?obs sdmx-dim:refPeriod <http://reference.data.gov.uk/id/gregorian-interval/2014-01-01T00:00:00/P3Y>}.
  OPTIONAL{
    ?obs <http://purl.org/linked-data/cube#measureType> <http://statistics.gov.scot/def/measure-properties/ratio>}.
  OPTIONAL{
    ?obs <http://purl.org/linked-data/cube#measureType> <http://statistics.gov.scot/def/measure-properties/percent>}.
  OPTIONAL{
    ?obs <http://purl.org/linked-data/cube#measureType> <http://statistics.gov.scot/def/measure-properties/mean>}.
  OPTIONAL{
    ?obs <http://purl.org/linked-data/cube#measureType> <http://statistics.gov.scot/def/measure-properties/rank>}
}
```

Fig. 2. The SPARQL query to retrieve compatible datasets [9]. The query searches for datasets that describe ratio, percent, mean, and rank statistical indicators about 2011 data zones and the time periods 2015, 2014-2015, or 2014-2016.



```
SELECT ?area ?price ?employmentdeprivation ?schoolattendancerate WHERE {
{
?a qb:dataSet <http://statistics.gov.scot/data/house-sales-prices>;
<http://purl.org/linked-data/sdmx/2009/dimension#refPeriod> <http://reference.data.gov.uk/id/year/2015>;
<http://purl.org/linked-data/sdmx/2009/dimension#refArea> ?area.
?area <http://publishmydata.com/def/ontology/foi/memberOf> <http://statistics.gov.scot/def/foi/collection/data-zones-2011> .
?a <http://statistics.gov.scot/def/measure-properties/mean> ?price .
}
OPTIONAL{
  ?u qb:dataSet <http://statistics.gov.scot/data/scottish-index-of-multiple-deprivation---employment-indicators>;
  <http://purl.org/linked-data/sdmx/2009/dimension#refPeriod> <http://reference.data.gov.uk/id/government-year/2014-2015>;
  <http://purl.org/linked-data/sdmx/2009/dimension#refArea> ?area;
  <http://statistics.gov.scot/def/measure-properties/ratio> ?employmentdeprivation.
}
OPTIONAL {
    ?f qb:dataSet <http://statistics.gov.scot/data/school-attendance-rate>;
    <http://purl.org/linked-data/sdmx/2009/dimension#refPeriod> <http://reference.data.gov.uk/id/government-year/2014-2015>;
    <http://purl.org/linked-data/sdmx/2009/dimension#refArea> ?area;
    <http://statistics.gov.scot/def/measure-properties/ratio> ?schoolattendancerate.
 }
}
```

Fig. 3. Part of the SPARQL query to retrieve data from the Scottish data portal. The query retrieves data for the 2015 house prices in all data zones, along with data from two compatible datasets describing (i) the ratio of employment deprivation in 2014-2015, and (ii) the ratio of school attendance in 2014-2015. The final SPARQL query can be found in [9].

2015. In the same way, in datasets with three-year intervals we selected 2014-2016. The final SPARQL query used to find all compatible datasets (step 4) is presented in Fig. 2 [9]. The query searches for datasets with (a) 2015, 2014-2015, or 2014-2016 year of reference, (b) 2011 data zones as the granularity level of the geography, and (c) ratio, percent, mean, or score (rank).

The SPARQL query of Fig. 2 resulted in 16 compatible datasets. In order to find the statistical indicators that could be used in the creation of a predictive model, we manually locked the different values of the dimensions of the datasets (step 5). For example, in the "Age of First Time Mothers" dataset[4] the "Age" dimension holds three values, i.e., (a) 19 years and under, (b) 35 years and over, and (c) all. If we lock the "Reference Period" dimension to "2014/15-2016/17′" and the "Reference Area" dimension to "2011 Data Zones", then we can result in three variables, one per different value of the "Age'" dimension. We, hence, resulted in 60 variables that come from 16 datasets of the Scottish data portal.

We then retrieved the data for the 60 selected variables (step 6). Towards this end, we submitted a new SPARQL query [9] to the Scottish data portal. Part of the query can be seen in Fig. 3. The query selects three statistical indicators for all 2011 data zones; (a) the prices of

[4] https://statistics.gov.scot/data/age-at-first-birth

houses in 2015, (b) the ratio of employment deprivation for the years 2014-2015, and (c) the ratio of school attendance during the years 2014-2015.

Ethics Statement

The present work did not involve human subjects, animals or information from social media platforms.

CRediT Author Statement

Areti Karamanou: Formal analysis, Data Curation, Software, Writing - Original Draft; **Evangelos Kalampokis:** Methodology, Conceptualization, Supervision, Writing - Review & Editing; **Konstantinos Tarabanis:** Supervision.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data Availability

Integrated Statistical Indicators from Scottish Linked Open Government Data (Original data) (Zenodo).

Acknowledgments

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

References

[1] A. Karamanou, E. Kalampokis, K. Tarabanis, Linked open government data to predict and explain house prices: the case of Scottish statistics portal, Big Data Res. 30 (2022) 100355 ISSN 2214-5796, doi:10.1016/j.bdr.2022.100355.

[2] S. Law, B. Paige, C. Russell, Take a look around: using street view and satellite images to estimate house prices, ACM Trans. Intell. Syst. Technol. (TIST) 10 (5) (2019) 1–19, doi:10.1145/3342240.

[3] V. Taecharungro, Google Maps amenities and condominium prices: Investigating the effects and relationships using machine learning, Habitat Int. 118 (2021) 102463, doi:10.1016/j.habitatint.2021.102463.

[4] E. Kalampokis, E. Tambouris, K. Tarabanis, A classification scheme for open government data: towards linking decentralized data, Int. J. Web Eng. Technol. 6 (3) (2011) 266–285, doi:10.1504/IJWET.2011.040725.

[5] M. Janssen, M. Hartog, R. Matheus, A. Yi Ding, A.G. Kuk, Will algorithms blind people? The effect of explainable AI and decision-makers' experience on AI-supported decision-making in government, Soc. Sci. Comput. Rev. (2021), doi:10.1177/0894439320980118.

[6] B. Ansari, M. Barati, E.G. Martin, Enhancing the usability and usefulness of open government data: a comprehensive review of the state of open government data visualization research, Gov. Inf. Q. 39 (1) (2022), doi:10.1016/j.giq.2021.101657.

[7] E. Kalampokis, D. Zeginis, K. Tarabanis, On modeling linked open statistical data, Web Semant. 55 (2019) 56–68, doi:10.1016/j.websem.2018.11.002.

[8] E. Kalampokis, A. Karamanou, K. Tarabanis, Interoperability conflicts in linked open statistical data, Information 10 (8) (2019) 249, doi:10.3390/info10080249.

[9] A. Karamanou, E. Kalampokis, & K. Tarabanis (2022). SPARQL queries for acquiring Integrated Statistical Indicators from Scottish Linked Open Government Data [Computer software]. https://doi.org/10.5281/zenodo.7304041.