Consumer
Data
Research
Centre

An ESRC Data
Investment

CONSUMER DATA RESEARCH CENTRE

Controlled Data
Project Proposal Form

For office use: Ref

Consumer
Data
Research
Centre

An ESRC Data
Investment

INTRODUCTION

This form should be completed by those wishing to access the Consumer Data Research Centre's (CDRC) controlled data collections and for those who wish to access both controlled and safeguarded data for the same project. You should consult the ***CDRC Data Service User Guide*** before completing the form. A CDRC data scientist will be assigned to you who can provide support in the application process.

Once submitted your proposal will be forwarded to the CDRC Research Approvals Group (RAG) for independent assessment. Projects will be assessed based on the criteria listed in Appendix 1. Please note this process generally takes 6-8 weeks.

Approval to access data will not be granted without evidence that the applicant has acquired ethical approval for the research through their institution, or supplied evidence that this is not applicable. For non-academic projects, where there is no approval process in place, the CDRC will assist the applicant in acquiring this.

PART A. PROJECT DETAILS

1. Contact Details

1.1. Lead applicant: Thehuan Hoang
Department and Institution: University College of London
Address: 279 Holloway Road, Flat A205, London UK N7 8FB
Email: the-huan.hoang.23@ucl.ac.uk
Telephone: 07587470653

1.2. Initial Proposal Form Reference: 2079

1.3. Title of Project

[Working title] Predicting trip attraction for travel demand planning based on POI typology and network centrality

2. Project Proposal Details

2.1. Abstract. Appropriate for a general audience. This may be used by the CDRC for reporting and publicity purposes as well as selecting RAG academic reviewers (max 150 words).

[Work in progress]

This paper examines the relationship between the volume of non-work trips to urban subzones and various factors such as the typology and density of points of interest (POIs), centrality within the transport network, and socio-demographic characteristics of the area. By doing so, we can uncover patterns of human movement in cities beyond the traditional commute-focused models. We employ a deep learning method (ANN) to train the predictive model, using large-scale mobility dataset for training and validation, and SHAP to exact feature importance tpo the target variable. By identifying the key attractors for non-work trips, we aim to inform urban planning and transportation strategies that cater to the diverse and evolving needs of urban dwellers.

2.2. Project description. A detailed description of the project, documenting the motivation, scope and aims of the intended research as well as the methods you will use in the proposed research. Please indicate whether you currently possess expertise in these methods, or whether the methodological expertise will be provided by a member of your research team (max 1500 words).

[Work in progress]

This paper examines the relationship between the volume of non-work trips to urban subzones and various factors such as the typology and density of points of interest (POIs), centrality within the transport network, and socio-demographic characteristics of the area. By doing so, we can uncover patterns of human movement in cities beyond the traditional commute-focused models. We employ a deep learning method (ANN) to train the predictive model, using large-scale mobility dataset for training and validation, and SHAP to exact feature importance tpo the target variable. By identifying the key attractors for non-work trips, we aim to inform urban planning and transportation strategies that cater to the diverse and evolving needs of urban dwellers.

2.3. Research Category:

2.3.1. Is this request for an Undergraduate, Masters project? U'grad ☐ Masters ☒

2.3.2. Is this request for a PhD project? PhD ☐

2.3.3. Is your project funded, commissioned or sponsored by a funding body or any other organisation? Yes ☐ No ☒ Funding application in progress ☐

Please include the name, postal and web address of your current or prospective funder, and your grant/project reference number (if applicable).

The research question was derived from a related proposed project by Transport for London regarding identifying activity centres in London. TfL does not provide funding for the research but may be interested in the findings.

2.4. Project Impact. Please describe the anticipated scientific and societal benefits of the project and the ways in which you intend to maximise those benefits (max 500 words).

Uncover human mobility patterns in a city without explicit OD matrices, or using more resource-intensive methods such as travel diary surveys, or snapshots such as censuses, with possible broader generalisability for other cities by using only open geospatial data

2.5. End Users. Who are the main end users of this research (academic research, central government, consultancy, industry, local government, NHS, public sector, third sector? List all that are applicable.)

Academic research (full dissertation), and public sector/third sector (presentation of findings)

2.6. Outputs and Publications. What are the intended outputs or publications arising from the use of these data? (For example, journal articles, PhD thesis, report for government department, policy documents for a local authority, White Papers, new software or other tools, etc.)

Academic research (full dissertation), and public sector/third sector (presentation of findings)

2.7. Research Team

Please list the names, affiliation and email addresses of all known members of your research team and all co-authors on any publication/presentation who will make use of the CDRC data. If you are a student please include your academic supervisor.

Applicants seeking access to controlled CDRC data are required to have safe researcher training, as offered by the Administrative Data Research Network (ADRN), HM Revenue and Customs (HMRC), Office of National Statistics (ONS) or the UK Data Service (UKDS) and maintain their accreditation throughout the period of access. If you have not previously completed such training, the CDRC will help you to access a course. Please state whether you are currently accredited.

Please attach the lead applicant’s CV.

Research Team *(add more rows if required)*

Title, Name	Department/ Institution	Institutional email address	Will be accessing controlled data: Yes/No	Will be accessing safeguarded data: Yes/No	Completed a safe researcher training course. If yes, please specify course and date of completion.
Thehuan Hoang	UCL	The-huan.hoang.23@ucl.ac.uk	Yes	Yes	N/A
Elsa Arcaute	UCL	e.arcaute@ucl.ac.uk	Yes	Yes	N/A

PART B. DATA REQUEST

1.1. Data Required. Please provide the following information for each dataset requested. If a variable required is not currently held by CDRC then a Data Case for Support[1] may be submitted. Please add more lines if required.

Data Partner	Data Set	Controlled/ Safeguarded	Access to Full Data Set requested or specific variables (list)	Geographic Extent	Temporal Extent
Geolytix	*Geolytix aggregated in-app location dataset*	*Controlled*	*Full Data Set*	*Greater London*	*Full (2021-2022)*

[1] If data not currently available please contact info@ucl.ac.uk for a CDRC Data Case for Support form.

1.2. **Justification** Why do you need access to this data and how will it be used in your project? Briefly explain why access to the proposed controlled/safeguarded data is needed for your research and why less detailed or disclosive versions of the data sources are not sufficient for your purposes. **Please provide, in the case of controlled data, a description of the data outputs you will want to take from the secure lab, what they will be used for and why they are safe.** Please note that outputs should be 'finished outputs' (see Appendix 3). An example of such an output is given below.

LSOA11CD	No. Customers	No. Male	No. Female	Average Basket Spend	Total Basket Spend 2016	Average Income
E01000054	30	15	15	23	70000	32000
E01000055	***	***	***	8	32000	27000
E01000056	12	###	###	2	10000	18000

*** - any count with a value less than 10 should be supressed
- any paired variable where the counts of less than 10 could be revealed should all be supressed

The dataset will serve as the ground-truth data to train a model that seeks to explain trip attraction into a specific urban area based on a variety of features to be explored in the research (POI typologies, centrality, employment density, etc.)

As the main purpose is to estimate total number of people attracted to a certain spatial unit from other units, the location data will be aggregated as such:

COUNT during most trafficked hour – COUNT during least trafficked hour of the day for each spatial unit across 7 days in one randomised week.

Moreover, there will be no need for unique identifiers or further socio-demographic variables to ensure maximum safety in dataset usage. Any count less than 10 to be suppressed

2. Data Linkage

2.1. **Data Linkage.** If your project will be linking more than one data source, describe which data sources will be linked and how the linkage will be done, including any specific variables that need to be linked (if known). If any of the data to be linked has identifying information as defined in the Data Protection Act 1998 or General Data Protection Regulations please provide details and if you are bringing this data with you please provide brief details of the data, source, method of collection and information about consent required to link (e.g. Data Owner or Controller's name, any documentation available or previous contact; if consent has already been achieved). Please note that no project that has the potential to re-identify individuals through data linkage will be approved.

n/a

PART C: ACCESS REQUIREMENTS

This section considers resource requirements. Once a project is approved and the secure facility site allocated you will be sent a site specific user guide detailing lab facilities and conditions. A summary of these can be found in Appendix 2.

1. Duration of Access

Progress of a CDRC project will be affected by the amount of time taken to secure access to datasets; you may need to be flexible with your research timetable. In order to help us assess the likelihood of your project being feasible please provide information about the following:

- Preferred project start date[2]: 06/06/2024
- Estimated duration of the project (remember to factor in time for peer review): 3 months
- Expected time you will spend on data analysis (in no. of days): 60 days
- Any known publication or other deadlines you are looking to meet: 23/08/2024 – Dissertation Submission Deadline

2. Access to Secure Service

2.1 Secure Facility Site. Please specify which CDRC secure lab you would prefer to undertake your data analysis. While we will attempt to facilitate your requests, we cannot guarantee that data will be made available at your preferred site.

Leeds secure lab ☐ UCL secure lab ☒ Trusted research environment/data safe haven ☐

2.2. Software. Please specify if you have any software requirements not already provided by the CDRC. A list of available software can be found in Appendix 2.

n/a

Do you have a licence for this software?

Yes ☐ No ☐

2.3. Commercial Software licence. Would you require a commercial licence for any CDRC owned software? If yes, please note you will be required to purchase this licence.

Yes ☐ No ☒

[2] RAG review process generally takes 6-8 weeks.

2.4. Hardware Configuration. If you think you will require significant computing power please specify your specific hardware configuration requirements here. Examples of a typical hardware requirement: 50GB storage, with 8GB of RAM and 4 processors (or quad-core).

n/a

3. Data Security Requirements for Data Being Brought to Centre

If you are bringing in external data, please be aware of the data security requirements stipulated by the licensor and list below. Please specify any specialist data security requirements that are required from the CDRC for the duration of the project. *Please note that these requirements may be stipulated in the licence terms and conditions of the original data.*

n/a

4. Technical Support Requirements

4.1. Please specify if you will have substantial storage requirements for any additional data that you would be importing into the secure lab.

Yes

4.2. Please specify any specialist support you anticipate requiring for the duration of the project. For example: training, support with data cleaning, GIS / mapping support. Please be as detailed as possible.

> ***Please note:*** *The CDRC has limited support services available and these are only offered at* ***CDRC Leeds****. We may need to discuss your requirements in more detail with you before we forward your application to the RAG along with our own assessment of the feasibility of your project in light of current resourcing. In some cases where your support exceeds CDRC resource capabilities, the RAG may return your application requesting you re-submit with support requirements factored into your own project’s resourcing. Projects which receive approval from the RAG will do so on the basis of the support requirements initially agreed. If additional support requirements emerge during the lifetime of the project, additional permission will be required from the RAG, which will necessarily impact the timeline of the project even if approved.*

n/a

5. Ethical Approval

5.1. Have you sought or are you seeking **ethical approval from an institutional ethical approvals panel** or any other appropriate body?

Yes ☒ No ☐

3.2. If *Yes*, please provide **evidence** of the status of the application or the outcome of the ruling issued. Please list what evidence you are enclosing below and return it as a separate attachment in PDF format when you return this application form. Feel free to add any comments below.

I will seek additional approval from UCL Ethical Research Committee to incorporate data from CRDC partners into my research

3.3. If *No*, please bring this to the attention of the CDRC as soon as possible so that routes for ethical approval may be discussed.

Click here to enter text.

3.4. If you believe that ethical approval is not required for your research, please provide justification for this below.

Click here to enter text.

6. Supporting Documentation

Please select which of the following supporting documentation are included with application:

☒ Lead applicant CV *(required)*

☐ Data Case for Support form *(if applicable)*

☐ Extra data *(if applicable)*

PART D: DECLARATION

By completing this declaration I hereby declare that the information included in this application form is true and correct to the best of my knowledge. I understand that any false or misleading information given by me in connection with my application may result in termination of the application process and/or other sanctions.

I also agree that I will be the single point of contact for progress updates and communication regarding the progress of the application.

I agree for my personal information to be used for the purposes of processing this application in accordance with the relevant data laws of the UK.

☒ I consent to my contact details being added to the CDRC contacts database so that the CDRC can send me notifications of CDRC related activities.

☒ I understand that forwarding this form by email constitutes an electronic signature.

☒ I understand that final approval for this project may require the additional submission of project approval forms.

Name: Thehuan Hoang

Date: 06/06/2024

APPENDIX 1: RAG CRITERIA FOR ASSESSMENT

The role of the RAG is to provide independent and transparent assessments of applications by researchers for access to data through both the CDRC Safeguarded and CDRC Secure services based on a set of standard evaluation criteria. RAG is independent to the CDRC and will include representation from the academic, big data, industrial sectors as well as the data partners concerned. For full Terms of Reference and membership see https://www.cdrc.ac.uk/data-services/using-our-data/

Criteria for Approval

- **Scientific advancement –** how the project has the potential to advance scientific knowledge, understanding and/or methods using consumer data;
- **Public good** – how the project has the potential to provide insight and/or solutions that could benefit society;
- **Privacy and ethics –** the potential privacy impacts or risks, and wider ethical considerations relating to the project
- **Project Design and Methods –** how the project will be conducted and who will be involved with a focus on demonstrating project feasibility.
- **Cost and resources issues –** what impact the project is likely to have on CDRC resources**,** including CDRC staff time and use of infrastructure, as well as any data acquisition costs. Resource requirements should be justified.

APPENDIX 2: CDRC SECURE DATA SERVICES & TOOLS AVAILABLE

JDI Research Lab

The CDRC have access to the Department of Security and Crime Science Jill Dando Institute Research Lab (JDI Research Lab) - a high-security computer facility where research using highly confidential data is undertaken. The JDI Research Lab hosts datasets on behalf of CDRC, providing access to these data for CDRC researchers, via workstations in a secure work area. To enter this secure work area users must attain clearance to meet the HMG Baseline Personnel Security Standard (BPSS).

The facility comprises a swipe card controlled windowless but air conditioned lab with workstations with twin 27” screens. These all run a virtualised windows operating system with each terminal providing the following software:

- ArcGIS
- CrimeStat
- MATLAB
- Office
- GeoDa
- MinGW
- NetLogo

- Notepad ++
- R
- Sophos
- QGIS
- Psycopg
- Python
- Enthought Canopy
- SQL 2014 Enterprise

University of Liverpool Secure Facility

The facility at the University of Liverpool comprises a swipe card controlled windowless but air conditioned lab with ten 27" all in one computers. These all run a virtualised windows operating system with each terminal providing connectivity to the CDRC secure database and to the following software:

- R and R Studio
- Python
- Anaconda
- Sublime Text Editor

On each desktop, there are some example scripts that can be used to access data, and help get you started on your research project.

University of Leeds Integrated Research Campus

The CDRC have access to the Integrated Research Campus (IRC) - a highly secure virtual research environment where research using highly confidential data is undertaken. The IRC hosts datasets on behalf of CDRC, providing access to these data for CDRC researchers, via thin clients in a number of safe rooms.

The safe rooms are air-conditioned and are accessed by an individually-assigned electronic fob, which will give access to the circulation corridors, the safe room area and the specific room assigned. Each room contains 2 thin clients with dual 24" monitors attached. A 1 person room and a 4 person room are also available. One of the thin clients in the 4 person room has dual 28" monitors attached. These all run a virtual desktop system which requires a two-factor authentication token to access. Either Windows or CentOS Linux can be provided as required. Each desktop can be pre-configured with the software required. The following software packages are generally available but please discuss with us if you want to use other software packages or programming environments:

- Office
- Notepad ++
- R (with RStudio)
- QGIS

- Python
- Java
- SQL Server Management Studio

All analysis and outputs are completed in the virtual research environment, and as such, **no data will enter or leave any of the CDRC facilities on the day of your visit**.

APPENDIX 3: STATISTICAL DISCLOSURE CONTROL - OUTPUT REQUIREMENTS

Outputs

Outputs requested should be 'finished outputs' i.e. the finished statistical analyses that you intend to present to the public. If requiring intermediate outputs for a particular reason e.g. to present initial findings then these may be considered if clearly presented and clearly explained. All outputs must be easy to read and interpret and how they are to be used explained.

Non-Disclosive Data

Taken from GSS/GSR Disclosure Control Guidance for Tables Produced from Surveys, October 2014

Social Surveys

- For the majority of surveys, outputs should be for large geographical areas, e.g. Country or Government Office Region, or in some cases Local Authority District (or equivalent). The level of geography should reflect survey design.
- Suppress or combine unsafe cells, i.e. where there are one or two units contributing to the cell.
- Where the sample size of a total or sub-total is one or two, suppress the whole row or column to which the total refers, including any zero cells (or combine neighbouring categories).
- In unweighted tables, cell suppression does not provide sufficient protection. Unsafe cells should only be combined with other cells.
- If unweighted sample base numbers are essential they should be conventionally rounded to base 10.
- Percentages may be released, provided it is not possible to deduce where only one or two units have contributed to the cell.
- Units may be individuals, families or households, communal establishments or any other unit whose confidentiality should be protected.

Subsamples

- For the majority of surveys, outputs should be for large geographical areas, e.g. regions, or in some cases Local Authority District (or equivalent). The level of geography should reflect survey design.

- Table design should be used to remove all unsafe cells, i.e. where there is one unit contributing to a cell. Variable categories should be combined or variables removed until only safe cells remain.
- Percentages may be released, provided it is not possible to deduce where only one unit has contributed to the cell.
- Units may be individuals, families or households or any other unit whose confidentiality should be protected.

Business Surveys: Magnitude tables

- A cell meeting both the following criteria is safe (otherwise the cell is unsafe):
    - there must be at least *n* enterprise groups in a cell (threshold rule)
    - the total of the cell minus the largest *m* reporting unit(s) must be greater than or equal to p% of the value of the largest reporting unit (p% rule)

Note that the values of the p% and minimum threshold parameter *n* and m should remain confidential, since knowledge of these values reduces the protection. The choice of p, n and m would usually be decided by the Responsible Statistician. Typical examples would be 2,3,4,5 (for n), and 2,3 (for m) and 5% 10% ,15%, 20% (for p).

- Table design should be used first to reduce the number of unsafe cells in a table where this is consistent with the main uses of the data.
- Cell suppression is the standard method used to protect tables with unsafe cells. The unsafe cells are suppressed, known as *primary suppressions*. Other cells must be suppressed to prevent the values of the unsafe cells being calculated by subtraction from the marginal totals of the table. These are known as *secondary suppressions*.
- Cell suppression does not generally provide protection from disclosure by differencing. Tables should be published using fixed categories to avoid disclosure by differencing. For example the same geographies and SIC codes should always be used.

Business Surveys: Count data

- Tables of count data are to be protected by redesign of the table to protect sensitive cells. If further protection is required other techniques such as controlled rounding to base 5 should be considered.
- Percentages or rates must be derived from rounded values.