Respondent

❮ | 3 | | Patricia Ternes Dallagnollo   ⌄ | **112:57**   ❯
Time to complete

Project Team & Funding Details

Proposals for intern projects are welcomed from all academic disciplines and partners, and interdisciplinarity is strongly encouraged: for example, a project may be co-supervised by academics in different Schools or an academic with an external partner.

1. Lead supervisor name and title *

> Dr Patricia Ternes Dallagnollo

2. Lead supervisor's University School/ Faculty/ Institute *

> Faculty of Environment, School of Geography

3. Lead supervisor's email address *

> p.ternesdallagnollo@leeds.ac.uk

4. Co-supervisor names, titles and Schools/Faculties/Institutes *

> Dr Victoria Houlden, School of Geography, Centre for Spatial Analysis and Policy Dr Nicolas Malleson, School of Geography, Centre for Spatial Analysis and Policy

5. Please specify the amount of funding you are providing for this project.
Please select all that apply.

N.B. the cost of a 6-month intern project is £18,500 (includes 6 months of intern salary, a £1,500 training, travel and subsistence budget for the intern, and £1,500 administrative costs). If you think you will require the use of the Leeds Analytics Secure Environment for Research (LASER; see section 5, Qs 36-37, for how to obtain costings for use of this platform) please indicate this here.
*

☑ £18,500 (6 month project cost)

☐ £37,000 (12 month project cost)

☐ Additional costs associated with the LASER secure data hosting platform (see section 5, Qs 36-37)

6. Funding source.

Please select your funding source from the list below. *

○ Self-funding from grant account

○ External partner funding

◉ Seeking Consumer Data Research Centre funding

○ Seeking Leeds Social Sciences Institute ESRC IAA funding

○ Other

7. If you selected 'other', please specify the nature of the funding below.

N/A

8. If self-funding, please provide the account code as well as your FRIO contact.

(Please write N/A if not relevant to you) *

N/A

9. If self-funding and your grant has not yet been set up, please provide an account code from your School which can underwrite the costs of the project until the grant account has been set up.

(Please write N/A if not relevant to you) *

N/A

External Partner Details

Intern projects should aim to collaborate with an external partner in support of LIDA's partnership-building efforts. The mechanism for this partnership can be through data sharing, funding provision or collaboration over a research question of particular interest to the partner.

A letter of support from your partner detailing any financial or in-kind contribution should be included with your application and there is the option to upload this letter to the form below.

10. External Partner *

Leeds City Council

11. Give brief details of the following:
i) An overview of the company/external partner and outline the current relationship with them. (Max 100 words.)
ii) The resources the external partner will contribute to the project (Max 200 words.).
iii) Whether an NDA is in place with the partner. *

i) Leeds City Council, the local authority of the City of Leeds, are partners on the broader application of the work (quantifying the ambient population in Leeds) and have supported the work through membership of the CDAS CDT and through collaboration on other, related, projects. ii) The data that will be used by this project are largely publicly available, so no direct resources are required from the Council. Instead, the role of the Council in the collaboration involves the co-creation of the research aims and ongoing discussions throughout the lifetime of the project to ensure that the research is relevant and will deliver real impact. iii) N/A (publicly-available data)

12. Do you have a letter of support from your external partner(s) detailing the level of cash/ in-kind contribution? This should be uploaded below. *

○ Yes

○ No

○ N/A

13. Please attach your letter of support from your external partner here by uploading file.

📄 Letter of Support Footfall 20210423_Patricia Ternes Dall.pdf (https://leeds365-m…

Project Details

14. Please select a desired project start date

(N.B. projects of 12 months duration must commence on Mon 4th October 2021) *

○ Mon 4th October 2021

○ Mon 4th April 2022

15. Please indicate the duration of your project

N.B. projects of 12 months' duration will have one intern working on the first 6 months of the project and a different intern working on the next 6 months of the project. This ensures that each intern gets exposure to two different domains, project teams and research questions in the course of their internship. *

○ 6 months

○ 12 months

16. Project Title *

Modelling ambient populations under different restriction schemes

17. Please provide a summary of the research which underpins the proposed work, including project/grant codes where possible (max. 300 words). *

This project builds on the prototype machine-learning model developed as part of a previous CDRC-funded internship[1] and will use the data insight gained through a forthcoming project[2]: 1. Predictive data analytics for urban footfall - CDRC-funded internship (2017) attempting to quantify and model ambient populations. 2. Measuring Ambient Populations during COVID in Leeds City Centre - CDRC-funded internship, starting April 2021. This project also aligns with a number of projects conducted at UoL in collaboration with partners: * Rapid Assistance in Modelling the Pandemic (RAMP): Urban Analytics. The RAMP:UA project will benefit from an improved understanding about how city centre usage varies under different lockdown scenarios as a means of informing its estimates of population mobility during lockdown. This project therefore offers an opportunity for the CDRC to exploit links with RAMP partners in government and academia. * Data Assimilation for Agent-Based Models (DUST) requires up-to-date footfall estimates to inform its real-time population mobility models so will benefit from the model produced here. * Analysing COVID-19 Mobility Responses through Passively Collected App Data - CDRC-funded internship. This LIDA project is analysing smart-phone generated mobility traces from Cuebiq data, and this data can inform this project. * Bringing the Social City to the Smart City (ESRC-Turing) will benefit from a greater understanding about how the dynamics of cities changes under government. Relevant background publications and research blogs include: * A. Whipp. et al. (2021) Estimates of the Ambient Population: Assessing the Utility of Conventional and Novel Data Sources. International Journal of Geo-Information. DOI:10.3390/ijgi10030131 * Ross, S. and Breckenridge, G. (2021) COVID for Christmas? Analysing patterns of Christmas mobility in the UK https://www.cdrc.ac.uk/covid-for-christmas-analysing-patterns-of-christmas-mobility-in-the-uk/ * N. Malleson et. al. (2017) Predictive data analytics for urban footfall http://surf.leeds.ac.uk/announce/2017/07/03/FootfallMachineLearning.html

18. Please provide a summary of the project and its main aims and objectives (max. 300 words).

This should be written in the form of a project abstract for a broad, non-expert audience as this may be used in literature about the Programme circulated with other project supervisory teams. Please indicate if any information is sensitive and not for general disclosure. *

The COVID-19 pandemic has had a huge impact on urban mobility, leading to two major questions: (i) how have cities changed during the pandemic? and (ii) which changes will remain as the pandemic subsides? To help answer these questions, this project will build on previous CDRC-funded work to create a spatio-temporal machine-learning model to estimate footfall around Leeds. It will consider the local urban configuration, external factors (like weather conditions) and, importantly, the impact of various mobility restriction measures. First, the model will be trained using data from the years before the pandemic, and then lockdown restriction conditions will be incorporated. The model will be able to estimate the overall change in footfall, as well as the heterogeneous impacts that restrictions will have on different local areas. In summary the project will: * Develop a suitable open-source footfall machine-learning model. * Calibrate and validate the model using footfall data from the CDRC (SmartStreetSensors) and from Leeds City Council (footfall cameras). * Develop a dashboard to present maps and related visual outputs to help policy makers to easily explore different scenarios The project partners, Leeds City Council, have a particular interest in better estimating how footfall in city-centre will vary as the pandemic, and related policies, evolve. Hence a specific case study will be designed around their immediate policy objectives at the start of the project, taking into account current conditions. Although based on Leeds, the work will be generalisable to other cities that have footfall estimates and could even be applied even where footfall data do not exist. Ultimately we aspire to attract further funding to create a nationwide footfall model, which would represent a great methodological advance as well as a contribution to outcomes like improving public health. This would also represent an extremely attractive outcome for the CDRC.

19. Methods and Software:

Please provide a summary of the methods the intern will be using and any software packages they will need providing with in order to carry out the work.

(max. 300 words) *

The compute and storage requirements are minimal for most of the project and can be conducted on medium-spec PC. For final experiments with larger-data sets, access to a high-performance scientific computing may be required. In summary, the intern will use: * Core data science analysis and visualisation techniques using Python or R for the introductory work. * Machine-learning for the modelling work using scikit-learn (or similar).

20. Please describe how your project will develop new data analytics methods/tools for deployment amongst researchers and partners.

(max. 300 words) *

The project will develop a machine learning model capable of simulating the footfall in Leeds. An existing CDRC-funded internship is conducting the necessary preliminary data collection and analysis, so it is envisioned that this project can begin on the modelling work

immediately. First, a simplified model will be used to identify the most important features in the footfall data, considering data prior to the pandemic and data at different stages of the pandemic. A Random Forest methodology is appropriate for this step. Then, the model will be refined and a recurrent neural network (RNN), like a Long short-term memory (LSTM), will be used to make short-term predictions about city occupation in different scenarios. These specific scenarios will be developed in collaboration with the project partners, taking account of the conditions in the city at the time that the project starts, and the immediate policy objectives of the Council. Given the rapidly-developing policy context that surrounds the pandemic, these specific policies and case studies have not been pre-supposed by the academic team.

21. Outputs, outcomes and future plans:

Provide a summary of the output(s) and outcome(s) that will arise from the project and future plans relating to this work.

It is important that you specify the ways in which the intern will be involved in these outputs/ outcomes.

(max. 300 words) *

The project proposes three outputs: * Output 1: A working machine learning model capable of simulating the footfall in Leeds, with associated visualisations, considering the urban configuration and other external factors – including different restriction measures. All code will be open source. * Output 2: A paper presenting the scientific contribution in a high-ranking computational geography journal such as Transactions in GIS. * Output 3: A data dashboard and associated blog post, aimed at policy makers, outlining the model outputs, maps, other visualisations, and the relevant software documentation. Different scenarios will be presented in order to help policy makers to keep the City Centre busy and safe. The intern will be directly involved in the 3 outputs. They will focus most heavily on Output 1. Output 2 will be jointly authored; the intern will benefit from support and guidance of experienced members of the supervision team. For Output 3, the intern will be guided in the use of tools and technologies to create interactive visualisations of the results, drawing on the expertise of the project partners to ensure that the outputs foster impact. The project will develop a model with the potential to be generalized beyond Leeds, for example, the model can be developed in different locations where data are available. This can be used to estimate future short-term footfall to support policy makers in different locations. In addition, by incorporating further data, the model can be expanded and improved towards a nationwide footfall model, which would represent a great methodological advance and offer significant potential impact as other cities across the country attempt to navigate the evolving pandemic. This new model will form the basis of two funding applications (details below).

22. Please explain how your project could lead to follow-on funding, e.g. future grant applications. *

The impact on policies and the possibility of generalization beyond Leeds offer opportunities for further development in a range of different directions. One immediate aim will be to develop the model in different locations where data are available, e.g. in cities such as Bristol, Newcastle, and Glasgow, and then expand the model to cover cities that do not have footfall sensors. This move towards a nationwide model is well aligned to two funding calls in particular. The first is the 'ESRC-Secondary Data Analysis Initiative'; ideal because the proposal will cross disciplinary boundaries – touching on fields such as social policy, social statistics, computing and methods – and will make use of existing ESRC-funded datasets. In addition, the PI is an early-career researcher, which is a career stage that the ESRC are specifically targeting with the call. Drafting of the SDAI application will begin in later parts of the project. The second proposal, which will focus on the methodological novelty (such a model has not yet been attempted) will be targeted at the EPSRC Postdoctoral Fellowship scheme in the Infrastructure and urban systems research area. A nationwide footfall model can contribute to outcomes like Productive and Healthy Nations, in particular to ambitions like improving prevention and public health and creating better living standards. The work is known to be within the remit of the call because the EPSRC are already funding a related Fellowship lead by Dr Nikolai Bode ('Model-driven construction of city-level pedestrian traffic maps over time') whom the investigators will collaborate with through joint fellowships at the Alan Turing Institute. In both applications, additional resources will be requested to fund further LIDA/CDRC interns to continue the work (two interns per application).

23. What is the anticipated impact of this project? (e.g. policy influence etc.)

(max. 300 words) *

This project offers the potential for considerable policy impact, initially in Leeds and later, through the creation of a national model, in further cities. In particular, the project aims to: * inform decision making and potentially influence policy in the face of the evolving pandemic, helping policy makers to balance the need for safe city centres with economic benefits for city-centre businesses; * identify lessons that can be learned regarding city-centre management during the pandemic and better understand how this has changed lifestyles; * deliver a better understanding about how people interact with their environment and which contextual factors – ranging from the impacts of the local environmental context to higher-level government policies – affect this; * create a model to predict future scenarios to the benefit of research and practice and informative visual outputs for interested researchers, citizens and policy makers (all code will be open source code for adaptation to future research)

Intern Specification

While each project will vary, interns will broadly be expected to: undertake quantitative analysis on core datasets; plan and manage their own research activity in collaboration with the project

investigators and other members of the research team; meet with clients and partners in support of LIDA's partnership-building efforts; participate in LIDA research meetings, events and seminars; generate outputs, including research papers, case studies and other materials for publication.

The following criteria will be used during the recruitment of the LIDA Data Scientist interns:

☐ Interns will have excellent all-round numerical skills that may have been gained in the context of quantitative social science, computing, mathematics/statistics, medical bioinformatics or population studies;
☐ Interns will have completed an Undergraduate or Masters degree in a relevant field;
☐ Interns will be passionate about using data to answer real-life questions using statistical or predictive techniques;
☐ Interns will have knowledge of a programming language and/or software for statistical computing;
☐ Interns will have good communication skills (verbal and written) and a willingness to work as part of a team.
☐ Interns may have experience of working with large unstructured data sets;
☐ Interns may have knowledge of data science, machine learning or data mining.

The following questions have been modelled on the data science competency framework established by the Edison Group (https://edison-project.eu/sites/edison-project.eu/files/filefield_paths/edison_cf-ds-release2-v08_0.pdf (https://edison-project.eu/sites/edison-project.eu/files/filefield_paths/edison_cf-ds-release2-v08_0.pdf)) to allow you to select the competencies it would be ideal for your intern to possess. This competency framework will also be used in the recruitment of the interns to enable us to appoint interns who match project needs. This is intended as a framework and guide to assist with the recruitment of suitable interns; we cannot guarantee the appointment of interns who match all of the competencies you select from the lists below.

N.B. Your project may not require many of the languages, tools, software product or skills competencies below, but we are providing breadth in order to allow for the ranges of projects which may be submitted across disciplines.

24. Please select all of the data analytics and statistical languages and tools from the list below which it would be useful for the intern to have some familiarity with. *

- ☑ R/ RStudio

- ☑ Python

- ☐ Matlab

- ☐ SAS

- ☐ Stata

- ☐ Julia

- ☐ SPSS

☐ WEKA

☐ KNIME

☐ Scala

☐ Orange

☐ Octave

☐ PHP

☐ Pig

☐ RapidMiner

☐ NLTK

☐ Ruby

25. Please select all of the databases and query languages from the list below
which it would be useful for the intern to have some familiarity with. *

☐ PostgreSQL

☑ MySQL

☐ Nettezza

☐ Oracle

☐ MS SQL Server

☐ No SQL

☐ MongoDB

☐ Cassandra

☐ Redis

☐ Accumulo

☐ Hive (Hadoop)

☐ UML

☐ ERWin

☐ DDL

26. Please select all of the data visualization/ reporting tools from the list below which it would be useful for the intern to have some familiarity with.

☑ Mathplotlib

☑ seaborn

☐ B3.js

☐ FusionCharts

☐ Chart.js

☐ Tableau

☐ Raphael

☐ Gephi

☐ Datawrapper

☐ Google Visualisation API

☐ Google Charts

☐ Flare

☐ ArcMap (ArcGIS)

☐ QGIS

☐ Excel VBA

☐ Power BI

☐ Qlik View

☐ Qlik Sense

☐ Linux

☑ R Markdown

27. Please select all of the Big Data Analytics tools from the list below which it would be useful for the intern to have some familiarity with.

- ☐ Spark
- ☐ MapReduce
- ☐ Hadoop
- ☐ Mahout
- ☐ Lucene
- ☐ NLTK
- ☐ Pregel
- ☐ DataLakes
- ☐ Flume
- ☐ Kafka
- ☐ Storm
- ☑ MS Azure
- ☐ HDInsight
- ☐ APS
- ☐ PDW

28. Please select all of the Development and Project Management Frameworks, platforms and tools from the list below which it would be useful for the intern to have some familiarity with.

- ☑ Python
- ☐ Java
- ☐ C/C++
- ☐ AJAX
- ☐ D3.js

- ☐ jQuery
- ☑ Anaconda/ Jupyter Notebook
- ☐ Jboss
- ☐ Vmware
- ☑ Git
- ☐ MS Azure
- ☐ Heroku
- ☐ Google Cloud Platform
- ☐ Flask
- ☐ Django

29. Please select all of the Mathematics, Probability and Statistics competencies from the list below which it would be useful for the intern to have some familiarity with. *

- ☐ Linear Algebra
- ☐ Vector calculus
- ☑ Regression
- ☐ Classification
- ☐ Discrete distribution
- ☐ Continuous distribution
- ☐ Normal distribution
- ☐ Joint probability distribution
- ☐ Data description
- ☑ Random samples
- ☑ Parameter estimation
- ☐ ANOVA

☐ Stochastic process

☑ Computer simulation

☐ Linear regression

☐ Correlation

☑ Multiple regression

☐ Nonparametric statistics

30. Please select all of the Machine Learning competencies from the list below which it would be useful for the intern to have some familiarity with.

☑ Validation

☑ Random Forests

☑ scikit-learn

☐ Handling missing values

☐ Cross-validation

☐ XGBoost

☐ Data Leakage

☐ Decision trees

☐ Naive Bayes classification

☑ Ordinary least square regression

☐ Logistic regression

☑ Time series

☐ Natural Language Processing

31. Please select all of the Deep Learning competencies from the list below which it would be useful for the intern to have some familiarity with.

☑ Neural networks

- ☐ Convolutional neural networks

- ☑ Recurrent neural network

- ☐ Keras

- ☐ PyTorch

- ☐ TensorFlow

- ☐ Deep neural network

- ☐ Stochastic gradient descent

- ☐ Overfitting and underfitting

- ☐ Dropout batch normalisation

- ☐ Binary classification

32. Desirable skills of the intern

- ☐ Experience working with external stakeholders

- ☑ Experience working with large unstructured data sets

33. Please indicate any other skills or experience it would be useful for your intern to have. For example, will it be important for the intern to have knowledge of the domain area of your project (e.g. crime or health analytics)?

> General data science and programming skills, either in R or Python, and experience with Machine Learning.

Data source(s), infrastructure and IG requirements of the project

Answers in this section will help us to understand the type of data you will be using, how this will be hosted and accessed and whether LIDA can support you with this.

Please note – proposals which already have access to their data, or are close to obtaining data access, will be seen as stronger.

34. Please provide details of the data source(s) that will be used during the course of the project. (max. 300 words).

Please note that you should already have access to the proposed data or be able to provide details that access is on track ready for the start of the project.

Please specify:

☐ any data sharing agreements that are already in place with third parties;
☐ details of the sensitivity of the data (e.g. anonymised/ strongly pseudonymised/ weakly pseudonymised/ personally identifiable information);
☐ how the data is currently accessed and stored;
☐ how the data will be stored during the course of the project. *

> The dependent variables, representing the urban configuration and other external factors, will be obtained from publicly available sources such as OpenStreetMap, the School of Environment Weather Station, school holiday databases, etc. The dependent variable, footfall, will be obtained from these sources: * SmartStreetSensors; a CDRC product for which access has already been granted. * Leeds City Council footfall cameras; publicly available. * Leeds City Council footfall sensors; 30 new WiFi-detection sensors that the Council are currently installing and will be available, via an API, in time for this project. A current CDRC-funded internship is focussed on accessing, analysing,and visualising these data -- typically a time-consuming process -- so it is envisioned that this project can begin modelling immediately. In addition, the project will also explore the opportunity to use smart-phone generated mobility traces from Cuebiq. An existing LIDA project is analysing these data with a view to estimating aggregate ambient population counts, if successful these data can inform this project as well.

35. Does your project have ethical approval? *

○ Yes

○ No

○ Pending

◉ Not required

36. Will you require use of LIDA's cloud based, highly secure data analytics IT platform LASER (Leeds Analytics Secure Environment for Research), for data storage and access?

    Please contact the LIDA Data Analytics Team to discuss your requirements and ask for costs of hosting your project on LASER: ircdst@leeds.ac.uk (mailto:ircdst@leeds.ac.uk). (Please note that any costs will be in addition to the cost of £18.5k per six month project).

    For more information on LASER: https://lida-data-analytics-team.github.io/laserdocs/ (https://lida-data-analytics-team.github.io/laserdocs/). *

    ◯ Yes

    ⦿ No

    ◯ Maybe

37. If you answered yes to the question above, please confirm you have spoken to the LIDA Data Analytics Team about your requirements and include the confirmed costs for using LASER below.

    Please write N/A if not planning to use LASER for data hosting and access for this project. *

    | N/A |

38. Please indicate whether your project will require data held by the Consumer Data Research Centre as detailed in its Data Store (https://data.cdrc.ac.uk/ (https://data.cdrc.ac.uk/)).

    Please complete an Initial Proposal Form (https://data.cdrc.ac.uk/using-our-data-services (https://data.cdrc.ac.uk/using-our-data-services)) in order to register your interest in accessing any safeguarded or controlled data sets hosted by the CDRC. *

    ⦿ Yes

    ◯ No

○ Maybe

39. If your project proposes to use data held by the Consumer Data Research Centre (CDRC), do you already have a user agreement in place with the CDRC? *

◉ Yes

○ No

○ N/A

40. Do you anticipate needing to use High Performance Computing (HPC) for your project? *

○ Yes

○ No

◉ Possibly

41. Will the intern working on your project require an NHS Honorary Contract? *

○ Yes

◉ No

○ Not yet known

Further Information

For further information about the LIDA Data Science Internship Programme please contact:

Kylie Norman, Programme Co-ordinator: k.r.norman@leeds.ac.uk
Or Paul Evans, LIDA Research & Innovation Development Manager: p.d.evans@leeds.ac.uk

Further information about the Data Science Internship Programme can be found at:
https://lida.leeds.ac.uk/study-training/data-science-internship-scheme/

(https://lida.leeds.ac.uk/study-training/data-science-internship-scheme/)

We will be sharing project proposals with the LIDA Intern Programme Academic Advisory Group who will be selecting projects for the 2021/22 Intern Programme in May. We hope to be able to let you know the outcome of your application after this.