Forecasting the restatement of house prices using machine learning

Yousef Gharib

1. Introduction

Mortgage Industry Advisory Corporation (MIAC) Analytics is an independent consultancy based in the United States and has been a leading provider of mortgage and financial services such as risk management and data auditing.

MIAC analytics have their own UK house price index based on actual property sold, as recorded by the Land Registry in the UK that is updated monthly; transactions are recorded very slowly by Land Registry meaning that it may take months to receive all the transactions that occurred in a single month. This issue has worsened since the start of COVID-19 resulting in the need of predicting the change in the aggregated house price (known as re-statements) as more volumes of transactions are recorded.

The formal explanation of the problem is, if $HPI(m, t_i)$ is the house price index value at time t_i calculated in month m, then generally for $n > 0$, $HPI(m, t_i) \neq HPI(m + n, t_i)$ due to the way data is obtained over time. If assumed that for $n = N$ large enough, we have $HPI(m + N, t_i) = N = HPI(m + n + k, t_i)$ for all $k > 0$ meaning we have all the transaction for month t_i in month $m + N$, and therefore know the index value at that point and have been approximating $HPI(m + N, t_i)$ using a subset of the full data.

Therefore, the research question this project aims to answer is:

How well can we predict future restatements given previous restatements and other macroeconomic data?

Through this we may be able to predict the restated values for lagged months where there is not enough data to calculate the aggregated price for that index.

Research objectives:

- *Data acquisition:* The data required consists of a mix of the index data which will be obtained from the MIAC database and macro-economic data which can be obtained from other sources such as Office for national statistics (ONS). The aim of this objective is to retrieve appropriate data for the next objectives.
- *Variation of restatements:* The time taken for restatements to converge can be calculated given the processed data; this may vary greatly based on factors such as location and property type. By finding the average time it takes for restatements to converge, we can use this information to decide how long of a window to consider when predicting a certain month in the time series.
- *Data preprocessing:* An exploratory analysis will be conducted to visualize properties of the data such as distribution and outliers. The data will also be processed through merging datasets and calculations of new features. This step is crucial, and the resulting dataset will be used for model training and evaluation.

- *Modelling:* Specific machine learning models are selected based on relevant literature and will be trained on the processed dataset. These models will be compared and used to predict the restatement for lagged months. This is the focus of the project and investigating how good of a prediction we can get using a machine learning model is the objective.
- *Evaluation:* The models will be assessed during the evaluation period using general machine learning metrics. This objective allows us to identify each model's performance and evaluate which are the best for this task.

This project is expected to have the following beneficiaries:

- The results from this project will directly benefit MIAC analytics as it may attract new customers if advertised as a product.
- Consumers of MIAC products may also benefit from this project as they may have exclusive access to predicted results and can make decisions based on the results.

2. Critical context

2.1. Time series forecasting

Time series forecasting is a popular economics problem revolving around predicting future observations by fitting a model to historical data. Classical models such as Auto regressive integrated moving average (ARIMA) and Vector autoregression (VAR) have generally been used for time series forecasting as they can be adapted to any time series problem easily and the only data required is the time series itself. The classical techniques tend to outperform machine learning techniques for time series; however, the multivariate predictability of machine learning techniques leaves a huge potential for machine learning models in predicting time series.

A plethora of machine learning models have been tested for time series forecasting and, in most cases, the optimal model depends on how many months are being forecasted; generally, models with a simpler architecture and input structure perform better for short-term forecasting. Support vector regressors tend to outperform other models for house price prediction [1,2], however LSTMs have shown to work well when forecasting for multiple locations [4].

2.2. Support Vector Regression (SVR)

Support vector machine is a popular machine learning algorithm generally used for classification tasks; this algorithm can also be adapted for regression tasks using similar principles and maintains its robustness as a machine learning algorithm. For classification tasks, a SVM constructs a hyper-

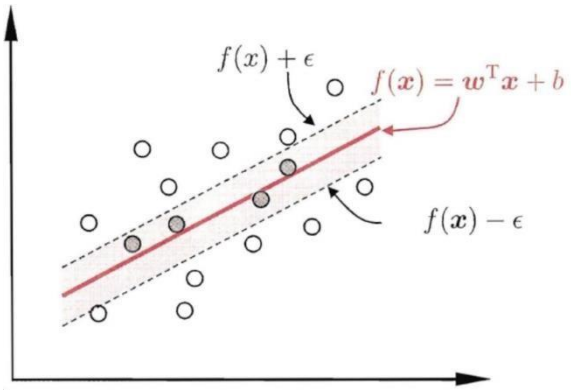


Figure 1: SVR [7]

plane to separate the classes in the dataset, the hyper-plane is constructed with the condition of maximizing the distance of the nearest data points for each class [6]; whereas for regression tasks, the aim is to find a function such that the difference between the actual data points and the function is within a threshold accuracy.

SVR has shown to perform excellently for time series forecasting [8] and has shown to outperform other popular machine learning models, specifically for short term prediction [1,2]. In addition to this, SVR can be highly effective with high dimensionality data through the application of kernel functions to transform the data, in turn allowing for easier optimization of loss functions. However, SVR works by solving constrained quadratic equations to minimize the loss function in which time can increase quadratically with the training set [9]. Combining hyper-parameter optimization with long SVR training times can end up taking a significant amount of time which may not be worth it.

2.3. Long short-term memory (LSTM)

Long short-term memory neural network is an artificial recurrent neural network that is mainly used in learning patterns of sequential data such as music composition, speech recognition and time series prediction. LSTMs work similarly to other recurrent neural networks in that the network loops over itself to allow information to be passed on at each step; the advantage of LSTMs over regular recurrent neural networks is that they are capable of learning long term dependencies through the cell state in the network structure, as shown by the top horizontal line in figure 2. The cell state stores additional information and is updated at each step.

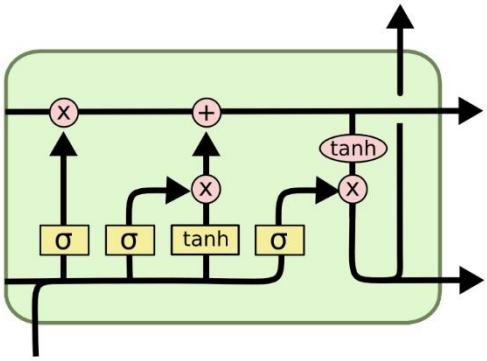


Figure 2: LSTM structure [13]

The ability to learn long-term dependencies is what makes LSTMs excellent at time series forecasting, since they can identify long term trends in the data. LSTM has become a popular model to test when using machine learning for house price prediction and has shown its ability to produce significant results [2,4]. Another advantage of LSTM is its adaptability through variants such as bidirectional and stacked LSTMs, as well as the ease of adjusting the model structure to allow for multiple outputs that may be based on geographic location as done by [4].

3. Approaches

3.1. Literature search and survey

The literature was sourced from Google scholar and other academic sources such as arXiv to validate the approaches chosen as well as give ideas on how the approach can be improved. An ongoing search for relevant literature will be carried out as seen necessary while working on objectives and writing the report.

3.2. Software and machine

The main software used for this project is Python as it provides all the necessary data preprocessing (Pandas), machine learning (Sklearn and Keras) and visualization tools (Seaborn and Matplotlib) to carry out the tasks set. Sequel query language (SQL) will be used for the data acquisition objective to obtain the necessary data.

The objectives will be carried out on a local computer. The use of cloud services will be decided on prior to model training and distributed learning will also be considered if model training times are longer than expected.

3.3. Data acquisition and preprocessing

The index data comprises of an aggregated price based for multiple locations and property types within the United Kingdom as well as dates ranging from January 1995 up to the specified index month; this data will be obtained from MIAC analytics database using SQL. Index data is available for each month from October 2018 up to March 2020 excluding lagged months and dates from when COVID-19 affects the index significantly. Data for each index month will be queried and appended on top of each other resulting in a large dataset consisting of all the necessary index data.

In addition to this, monthly macro-economic data will be obtained from multiple sources to help with the prediction, it is important to investigate if the macroeconomic data is lagged to ensure the dates for all the data match, this avoids issues where we are unable to predict due to missing data points. The initial macro-economic variables chosen are Gross domestic product (GDP), Unemployment rate and Interest rate; these were chosen as they were some of the most common macro-economic variables used to predict house prices, as seen by [1,2]. Extra macro-economic data such as inflation rate may be considered after testing to try and improve the models' performance. Annual macro-economic data may also be considered in addition to monthly data to reduce the variance of macro-economic data and incorporate overall trends instead of monthly change.

Once the data has been obtained and merged, a new column will be created to calculate the restatement as mathematically defined in equation 1.

$m = \text{Index month}$

$i = \text{Date in specific index month}$

$$Restatement_{m,i} = 100 \times \left(\frac{Price_{m+1,i}}{Price_{m,i}} - 1\right) \quad (1)$$

The final preprocessing step is to add a window of previous months of length *n*, where *n* is the average time taken for restatements to converge. This step is done by creating a new column in the dataset and shifting the restatements down by 1 for each column, as shown by table 1.

Table 1: Example of sliding window method for n = 4

Restatement	Restatement_-1	Restatement_-2	Restatement_-3	Restatement_-4
0.5	*NaN*	*NaN*	*NaN*	*NaN*
0.3	0.5	*NaN*	*NaN*	*NaN*
0.11	0.3	0.5	*NaN*	*NaN*
0.43	0.11	0.3	0.5	*NaN*
0.314159	0.43	0.11	0.3	0.5

This method incorporates structure into the model training by using previous *n* months restatements as predictor variables and has been used in many cases for time series forecasting using machine learning [3].

3.4. Variation of restatements

The final dataset of all the appended index months will be used to estimate the average time taken for restatements to converge, this gives us an idea of the number of months it takes for prices to stabilize in the index and informs us of the expected time taken for the total volume of data obtained each month to stabilize since we only expect the price to change if more transactions are obtained each month.

Based on the data, a new table with new variables will be created and defined to calculate the average time taken for restatements to converge along with other features. The time to convergence is found by investigating at what point the restated price (change in price) is not significant, defined as $Restatement_{m,i} < e$ where e is a threshold chosen and $e > 0$. Any number may be chosen as this threshold; restatements may vary greatly based on location and date therefore implementing a threshold that adapts to the location may result in a more reliable average time to convergence, an example of such implementation is using a threshold that is a combination of mean and standard deviation of the restatement values.

The average time taken will be used as the variable *n* to determine the window size to include in model training, doing this incorporates sequential data into the model as required for time series forecasting.

3.5. Modelling and testing

This project uses similar implementation to relevant literature on predicting house prices, however since we are predicting the restatements, the data results in being a time series within a time series; meaning the way the data is fed to the models is different compared to regular time series problems. Figure 3 displays how normal time series data is fed to a machine learning model. For a given time series, months used as training data to predict future months. Figure 4 shows how the time series data will be fed to the model for this project. The data is comprised of multiple datasets representing each index month, within each index month is the date and aggregated price of transactions that occurred at that point in time. Therefore, the time series is comprised of the restated price calculated between index months, for a specific date.

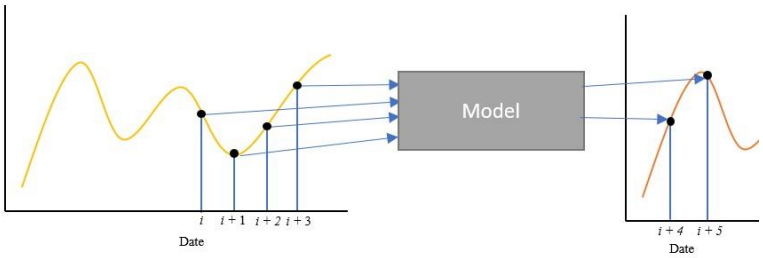


Figure 3: How normal time series data is fed to a model.

While the time series is slightly different, it is expected that all methods found in relevant literature will be applicable and provide similar results in this case.

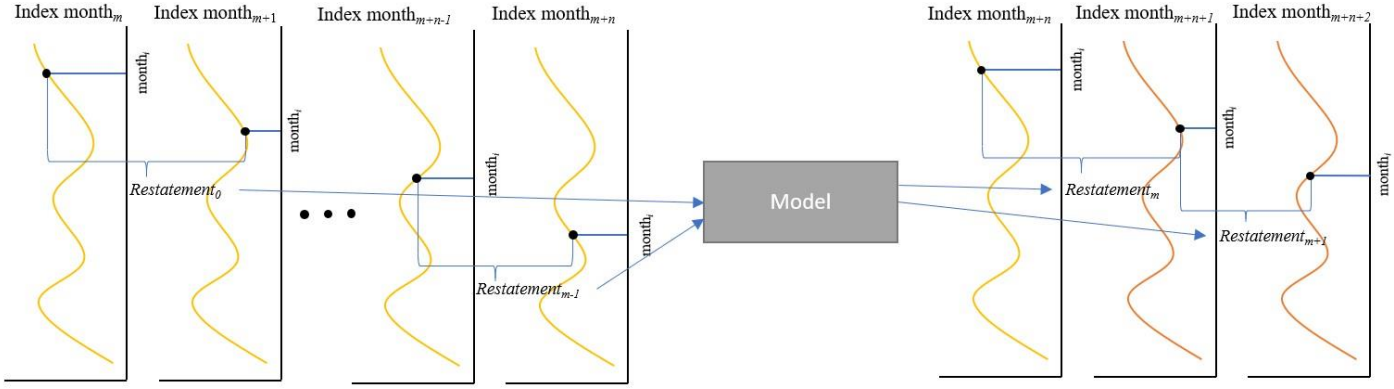


Figure 4: How time series data will be used

The initial machine learning models are chosen based on their performance in relevant literature. Since the focus is short term forecasting, the initial models chosen are support vector regression (SVR) and long short-term memory neural networks (LSTM) as they have shown to perform well for short term house price forecasting [1,2,4]. Other models may be considered such as Random forests as it is an ensemble method (and we are working with multiple locations and property

types), in addition to multi-layer perceptron and LSTM variants such as Bidirectional LSTM as it has potential to outperform other mentioned models [5].

Model hyperparameters will be optimized using a grid-search, if it is found that model training times are too long due to this, other hyperparameter optimization methods such as Bayesian optimization may be implemented instead. For a fair comparison, all models will be optimized and tested using the same training and test set. Given the timeframe it is expected that sufficient time will be available to test models in addition to SVR and LSTM, but the initial objective is to test and evaluate those two models. Model implementation and hyperparameter optimization will be carried out using prebuilt python libraries, some models such as bidirectional LSTM may be tougher to implement as code for the model may need to be obtained from online resources, however this will be investigated if the decision to test the model is made.

3.6. Results and evaluation

Standard regression metrics such as mean squared error and mean absolute percentage error will be used to evaluate the models chosen. Model metrics will also be considered where appropriate, such as loss criterion for neural networks.

A random walk with drift will be built to use as a benchmark model to compare the trained models with [1,2]; in addition to using the random walk as a comparison, it helps us comprehend the predictability of our time series forecasts problem.

3.7. Report

The report will be written in parallel with the work carried out to allow for greater detail to be noted. The period will be set before the deadline that will be dedicated to completing the report to allow for feedback and leave enough time to address areas that can be improved. To avoid falling behind schedule, some extra time is allocated to objectives to take unexpected issues into account and avoid falling behind schedule.

3.8. Project meeting

A meeting is scheduled once every 2 weeks with the project supervisor to discuss progress made and possible feedback and advice.

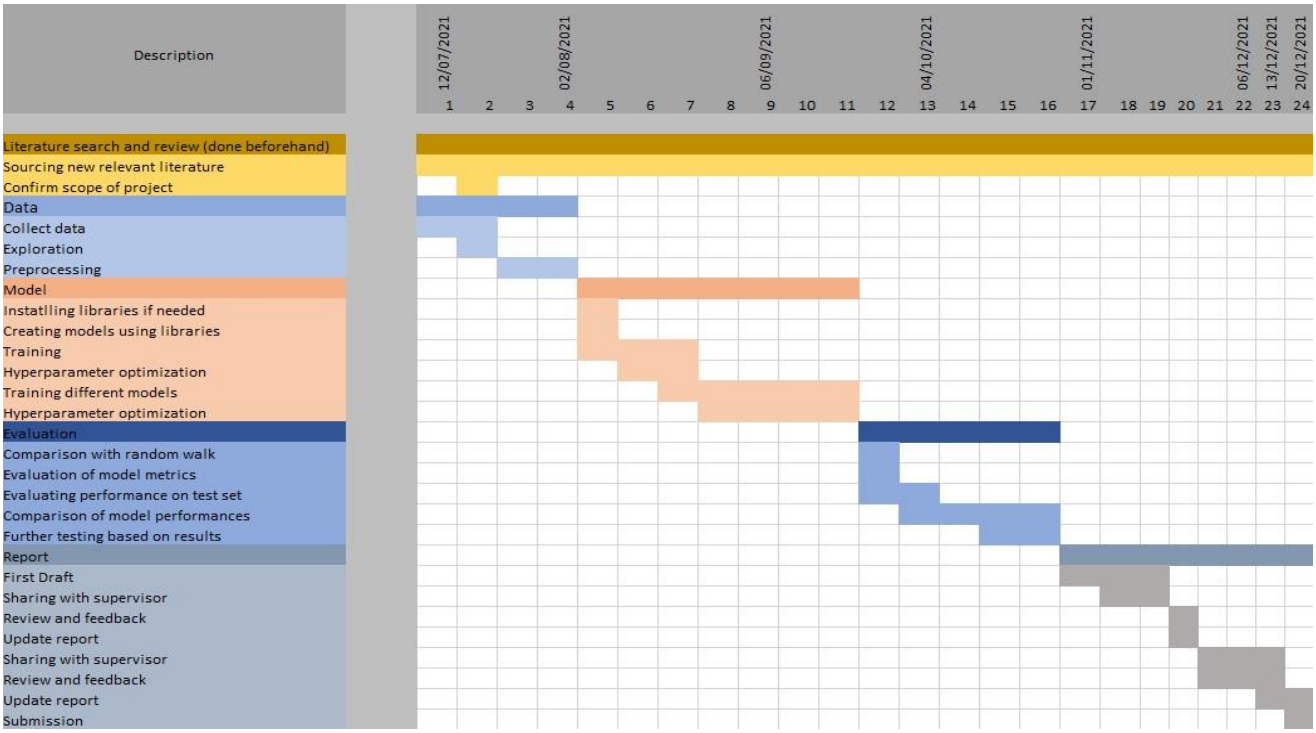
4. Risks

Description	Likelihood (1-3)	Consequence (1-5)	Impact (L x C)	Mitigation
Unable to achieve key milestones	1	5	5	Carry out more research and discuss with supervisor
Model training taking a long time	3	3	9	Adjust training method or use cloud services
Deliverables affected due to illness	2	2	4	Extra time will be allocated between milestones to take external factors into account

Implementation too difficult	1	4	4	Code carrying out similar tasks will be sourced and used as a reference
Loss of motivation	1	4	4	Discuss with supervisor
Code accidentally lost or deleted	1	5	5	Code regularly pushed on version control software e.g., GitHub

5. Workplan

Phase	Description	Outcome
Literature search and review	• Sourcing new relevant literature • Confirm scope of project	Updated Approach
Data acquisition and pre-processing	• Collect data • Exploration to make sure the dataset format is understood • Data processing	Data ready to be used for model training
Variation of restatements	• Create new variables and calculate the average time taken for restatements to converge	Use the average time as the window in model training
Modelling and evaluation	• Installing libraries if needed • Using libraries to create models • Train models on data • Hyperparameter optimization • Training different models	Models will be able to predict lagged months and provide all the necessary results for evaluation
Evaluation	• Comparison with random walk • Evaluation of model metrics and testing on test set • Comparison of model performance • Ideas for further testing	Provides required information to answer research question and written analysis.
Report writing	• Cycle of drafting, getting feedback from supervisor and addressing points until submission	Submit final report



References

[1] Plakandaras, V., Gupta, R., Gogas, P. and Papadimitriou, T. (2014). Forecasting the U.S. Real House Price Index. *SSRN Electronic Journal*.

[2] Milunovich G. (2019). Forecasting Australian Real House Price Index: A comparison of Time series and Machine learning methods.

[3] Hota, H., Handa, R. and Shrivas, A. (2017). Time Series Data Prediction Using Sliding Window Based RBF Neural Network. *International Journal of Computational Intelligence Research*, [online] 13(5), pp.1145–1156 [4]

Chen, X., Wei, L. and Xu, J. (n.d.). *House Price Prediction Using LSTM*.

[5] Kutlualp, A. (n.d.). *Classical Machine Learning vs. Deep Learning Second Elizabethan Age Financial Portraiture PostEurope: Forecasting the GBP/USD Exchange Rate in the Era of Brexit*.

[6] V.Kecman (2013). Support Vector Machines: Theory and Applications.

[7] Programmersought.com. (2018). *[Machine Learning] Regression--Support Vector Regression (SVR) - Programmer Sought*.

[8] Debasish Basak, Srimanta Pal and Dipak Chandra Patranabis (2007). *Support Vector Regression*. [online] ResearchGate. Available at: https://www.researchgate.net/publication/228537532_Support_Vector_Regression [Accessed 4 Aug. 2021].

[9] Bottou, L. and Lin, C.-J. (n.d.). *Support Vector Machine Solvers Support Vector Machine Solvers*. [online] . Available at: https://leon.bottou.org/publications/pdf/lin-2006.pdf [Accessed 4 Aug. 2021].

Research Ethics Review Form: BSc, MSc and MA Projects

Computer Science Research Ethics Committee (CSREC) http://www.city.ac.uk/department-computer-science/research-ethics

Undergraduate and postgraduate students undertaking their final project in the Department of Computer Science are required to consider the ethics of their project work and to ensure that it complies with research ethics guidelines. In some cases, a project will need approval from an ethics committee before it can proceed. Usually, but not always, this will be because the student is involving other people ("participants") in the project.

In order to ensure that appropriate consideration is given to ethical issues, all students must complete this form and attach it to their project proposal document. There are two parts:

PART A: Ethics Checklist. All students must complete this part. The checklist identifies whether the project requires ethical approval and, if so, where to apply for approval.

PART B: Ethics Proportionate Review Form. Students who have answered "no" to all questions in A1, A2 and A3 and "yes" to question 4 in A4 in the ethics checklist must complete this part. The project supervisor has delegated authority to provide approval in such cases that are considered to involve MINIMAL risk.
The approval may be ***provisional*** – *identifying the planned research as* likely to involve MINIMAL RISK. In such cases you must additionally seek ***full approval*** from the supervisor as the project progresses and details are established. ***Full approval*** must be acquired in writing, before beginning the planned research.

A.1 If you answer YES to any of the questions in this block, you must apply to an appropriate external ethics committee for approval and log this approval as an External Application through Research Ethics Online - https://ethics.city.ac.uk/		*Delete as appropriate*
1.1	Does your research require approval from the National Research Ethics Service (NRES)?	**NO**
1.2	Will you recruit participants who fall under the auspices of the Mental Capacity Act?	**NO**
1.3	Will you recruit any participants who are currently under the auspices of the Criminal Justice System, for example, but not limited to, people on remand, prisoners and those on probation?	**NO**
A.2 If you answer YES to any of the questions in this block, then unless you are applying to an external ethics committee, you must apply for approval from the Senate Research Ethics Committee (SREC) through Research Ethics Online - https://ethics.city.ac.uk/		*Delete as appropriate*

2.1	Does your research involve participants who are unable to give informed consent?	**NO**

2.2	Is there a risk that your research might lead to disclosures from participants concerning their involvement in illegal activities?	**NO**
2.3	Is there a risk that obscene and or illegal material may need to be accessed for your research study (including online content and other material)?	**NO**
2.4	Does your project involve participants disclosing information about special category or sensitive subjects?	**NO**
2.5	Does your research involve you travelling to another country outside of the UK, where the Foreign & Commonwealth Office has issued a travel warning that affects the area in which you will study?	**NO**
2.6	Does your research involve invasive or intrusive procedures?	**NO**
2.7	Does your research involve animals?	**NO**
2.8	Does your research involve the administration of drugs, placebos or other substances to study participants?	**NO**
A.3 If you answer YES to any of the questions in this block, then unless you are applying to an external ethics committee or the SREC, you must apply for approval from the Computer Science Research Ethics Committee (CSREC) through Research Ethics Online - https://ethics.city.ac.uk/ **Depending on the level of risk associated with your application, it may be referred to the Senate Research Ethics Committee.**		*Delete as appropriate*
3.1	Does your research involve participants who are under the age of 18?	**NO**
3.2	Does your research involve adults who are vulnerable because of their social, psychological or medical circumstances (vulnerable adults)?	**NO**
3.3	Are participants recruited because they are staff or students of City, University of London?	**NO**
3.4	Does your research involve intentional deception of participants?	**NO**
3.5	Does your research involve participants taking part without their informed consent?	**NO**

3.5	Is the risk posed to participants greater than that in normal working life?	**NO**
3.7	Is the risk posed to you, the researcher(s), greater than that in normal working life?	**NO**
A.4 If you answer YES to the following question and your answers to all other questions in sections A1, A2 and A3 are NO, then your project is deemed to be of MINIMAL RISK.		*Delete as appropriate*
If this is the case, then you can apply for approval through your supervisor under PROPORTIONATE REVIEW. You do so by completing PART B of this form. **If you have answered NO to all questions on this form, then your project does not require ethical approval. You should submit and retain this form as evidence of this.**		
4	Does your project involve human participants or their identifiable personal data?	**NO**