Technical Appendix

Analysis of UK Vacancy Time Series Data Across Vintages

A. Objective

This appendix delineates the technical implementation of a time series analysis conducted on historical UK vacancy estimates published by the Office for National Statistics (ONS). The study investigates how vacancy estimates evolve across successive data vintages and applies statistical modelling to forecast future vacancy levels. The analysis emphasises reproducibility, modular design, and interpretability.

B. Data Acquisition and Automation

A class-based Python procedure was developed to automate the retrieval and preprocessing of historical CSV files from the ONS archive. The class "VacanciesAcquisition" encompassed all functions related to data acquisition, automation, and local storage preparation.

- **Libraries:**
  - requests – for HTTP requests,
  - os – for file system operations,
  - time – for rate limiting,
  - logging – for structured logging and error reporting,
  - pandas – for data manipulation.

- **Key Functionality:**
  - **Directory Creation for Local Storage:** A dedicated folder (ONS_UK_Vacancies) is created to store downloaded files. The implementation includes controls to avoid overwriting or duplication, ensuring safe re-execution.
  - **Dynamic URL Construction for Each Vintage (Release Date):** Parameterised download links are generated for each vintage using a consistent URL pattern. This allows scalable access to historical releases without manual intervention.
  - **Download Execution and Validation:** HTTP GET requests are sent to the ONS server. Response status codes are validated, and content is written to disk in binary format. Logging records successful and failed downloads.
  - **Error Management:** Exception handling is implemented to capture and report failed downloads without interrupting the overall process. A delay of 3 seconds between requests ensures compliant server access.

C. Data Preparation and Structuring

Each downloaded CSV file comprises both metadata and time series observations. To extract, validate, and structure the data, a second class “VacanciesStructuring” was defined, containing three dedicated Python function applied sequentially.

- **Libraries:**
  - pandas (data manipulation and structuring),
  - re (regular expression parsing).

- **Key Functionality:**
  - **Vintage Extraction** - extract_vintage_date(filepath): This function parses the release date from the metadata header using regular expressions. The extracted string is converted into a standardised datetime object to represent the vintage of the dataset.
  - **Monthly Series Parsing** - parse_monthly_series(filepath): This function filters the time series content to retain only rows corresponding to monthly observations, identified by the pattern YYYY MMM. It converts date strings into datetime objects and vacancy values into numeric format. Each observation is annotated with its corresponding vintage.
  - **Data Consolidation** - consolidate_all_data(): This function aggregates all valid monthly datasets into a unified pandas DataFrame. The resulting structure contains three fields:
    i. Observation date: the calendar month to which the vacancy estimate pertains,
    ii. Vacancy estimate: the reported number of vacancies (in thousands),
    iii. Vintage date: the release date of the estimate, indicating its version.

The consolidated dataset is exported as ONS_UK_Vacancies_Consolidated.csv using UTF-8 encoding and semicolon delimiters.

D. Visualisation of Revisions

To examine how UK vacancy estimates evolve across successive data releases, two complementary visualisation approaches were employed:

Static Chart (Plotly Express):

A focused line chart was produced to illustrate how the estimate for February 2024 changed across vintages. This visualisation highlights the revision path for a single month, enabling the assessment of how initial figures are refined over time.

- The chart is designed to identify whether early values tend to be overestimated or underestimated, and how quickly these stabilise.
- A stable line suggests minor revisions, whereas fluctuations may indicate underlying uncertainty or methodological bias.

Interactive Chart with Dropdown Functionality (Plotly Graph Objects):

A dynamic chart was developed to allow exploration of revisions across all available months. Each trace represents a specific month's vacancy estimates across its vintage history.

- A dropdown functionality enables users to select any month, updating both the visible data and the chart title accordingly.
- To ensure chronological accuracy, the dropdown options are sorted by actual observation date rather than alphabetically, thereby preserving the temporal flow of the data and avoiding misleading groupings.

Although the exercise required visualising revisions for a single month, additional months were included to provide broader context and to examine the consistency of revision patterns across multiple months. This extended view supports a more robust interpretation of the data and strengthens the foundation for subsequent forecasting.

February 2024 Vacancy Revisions - Key Observations:

- Initial Estimate (~916k): The first release shows a relatively high vacancy level. This likely reflects early, incomplete data, possibly optimistic or based on preliminary surveys.
- Early Drop (~913k): A slight downward revision follows quickly, suggesting the initial figure was adjusted as more data came in. This is typical of early vintage behaviour: fast corrections within the first few months.
- Stabilisation (~919k): From mid-2024 to early 2025, the estimate remains approximately at 919k. This plateau suggests a period of confidence in the estimate; either no new data arrived to challenge it, or the methodology remained consistent.
- Sharp Final Drop (~905k): Around mid-2025, the estimate declines significantly and remains low. This late correction could be due to benchmarking updates, inclusion of lagging data sources, or methodological reclassification. Its persistence at approximately 905k across multiple vintages suggests that this figure is now regarded as the definitive estimate.

Revision path:

The sequence of revisions typically begins with early volatility, transitions into a period of apparent stability, and culminates in a late-stage structural adjustment. These revisions are not random; rather, they appear to follow a delayed maturation trajectory shaped by the data collection and refinement process.

This observation has direct implications for the forecasting model we intend to develop. While initial estimates may serve as useful indicators for short-term

analysis, they lack reliability for long-term inference. Accordingly, finalised values should form the basis of any robust forecasting framework.

E. Forecasting

Forecast Model:

To forecast future UK vacancy levels, an ARIMA(1,1,1) model was implemented using the finalised time series; constructed by selecting the latest available vintage for each calendar month. This ensures that the input data reflects the most stable and accurate estimates, avoiding distortions from preliminary revisions.

Steps:

- **Data Preparation:**
  Extracted the latest vintage for each reference month to construct a finalised time series.
- **Stationarity Testing:**
  Augmented Dickey-Fuller test yielded a p-value of 0.347, indicating non-stationarity. Consequently, first-order differencing was applied.
- **Model Identification:**
  ACF and PACF plots of the differenced series revealed a strong autoregressive signal at lag 1. Thereby, an ARIMA(1,1,1) model was selected.
- **Model Estimation:**
  The ARIMA(1,1,1) model produced statistically significant coefficients and demonstrated superior performance compared to more basic alternatives (e.g., ARIMA(1,1,0)), based on AIC.
  - AR(1) = 0.6734 ($p<0.001$)
  - MA(1) = 0.1953 ($p=0.01$)
  - AIC = 2418.056
- **Residual Diagnostics:**
  - **Ljung-Box Test (lag 10):** $p = 0.998$ → no autocorrelation in residuals.
  - **Jarque-Bera Test:** $p = 0.00$ → residuals not normally distributed.
  - **Heteroskedasticity Test:** $p = 0.00$ → residual variance not constant.

  Despite non-normality and heteroskedasticity, the model is statistically adequate due to the large sample size (n = 290). Confidence intervals remain valid under the Central Limit Theorem.
- **Forecasting:**
  Six-month forecasts were generated using get_forecast() with confidence intervals derived from the model's estimated residual variance. These intervals provide a credible range for future values.
- **Visualisation:**
  Forecasts are presented alongside historical data, with shaded confidence bands. The output is saved as vacancy_forecast.html.

Forecast Performance:

Forecast performance should be assessed by comparing predicted values against the finalised estimates for each month. Since vacancy data undergoes systematic revisions across vintages, early releases may differ significantly from their final value. Therefore, evaluation must account for this maturation process:

- **Use finalised values as reference standard:** Avoid comparing forecasts to early vintages.
- **Apply standard metrics:** Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and Mean Absolute Percentage Error (MAPE). These metrics provide complementary perspectives on forecast accuracy:
    - MAE quantifies average absolute error in the same units as the data (thousands of vacancies).
    - RMSE penalises larger errors more heavily, useful for detecting volatility.
    - MAPE expresses error as a percentage, allowing scale-independent comparison across time.

In addition to numerical metrics, visual diagnostics are essential:

- **Forecast Error Plots:** Residuals plotted over time can reveal structural shifts, seasonal bias, or periods of instability.
- **Rolling Accuracy:** As new vintages become available, forecast accuracy should be re-evaluated to assess model robustness under data maturation.

To deepen the analysis and improve model reliability, the following extensions are proposed:

- **Revision Modelling:**
  Vacancy estimates undergo systematic revisions across vintages. A natural extension is to model the revision path itself i.e.; how early estimates evolve toward their finalised values. This could be achieved using:
    - State-space models to capture latent dynamics and measurement error.
    - Hierarchical time series models to jointly model multiple vintages per month.
    - Kalman filters to update estimates as new vintages are released.
- **Bayesian Uncertainty Quantification:**
  Traditional ARIMA models provide confidence intervals based on residual variance. A Bayesian approach would allow full posterior distributions over forecasts and revision paths, enabling:
    - Credible intervals that reflect parameter uncertainty.
    - Probabilistic statements about future values and revision likelihoods.
    - Integration of prior knowledge, for instance historical revision behaviour.
- **Non-Linear Modelling:**
  While ARIMA captures linear dependencies, real-world revisions may follow non-linear patterns. Machine learning models could be used to learn these relationships:

- Gradient Boosting Machines (GBM),
- Random Forests,
- Neural Networks

These models could be trained to predict both vacancy levels and expected revision magnitude.

- **External Indicator Integration:**
  Vacancy trends are influenced by broader economic conditions. Incorporating external variables could improve forecast accuracy:
  - GDP growth rates,
  - Unemployment rates,
  - Business sentiment indices,
  - Sectoral employment data.

  These indicators could be used as exogenous regressors in ARIMAX models.