



STUDENTS AS
CO-CREATORS

A Disciplinary Research Collaboration

*Analysing and characterising flight plans and
trajectories*

Oscar Denche

School of Computer Science and Engineering

Luis Delgado and Michal Weiszer

School of Architecture and Cities

Academic Year 2025-2026

1. Executive Summary

This project developed a data driven method for characterising realistic flight route alternatives and their associated costs, intended as an input for Mercury, an air transport mobility simulator developed at the University of Westminster. Mercury already supports multiple flight plans per origin-destination (O-D) pair; the practical constraint is that the flight-plan data feeding it is several years old, and newer O-D pairs often have no derived alternatives at all. This project builds a pipeline that characterises route alternatives and their costs directly from historical flight-plan data, as a step towards keeping that input current.

Using EUROCONTROL flight-plan data from September 2023 (898,744 matched flights across the FIR-distance and flight datasets), a three-layer clustering pipeline was built: flights were first grouped by O-D pair, then split by which flight information regions (FIRs) they crossed (a binary signature), then further split by the actual distance flown within each FIR for that signature group. A deterministic cost model, built from EUROCONTROL CRCO unit rates, Nav Canada charges, and a fuel-burn formula, was attached to each resulting cluster.

The pipeline was developed and validated on three training O-D pairs (EGLL-KJFK, LEBL-LEPA, LPPT-EDDB), which produced 16, 1 and 1 clusters respectively, then tested on a held-out O-D pair (EGLL-LGAV) never used during development. On this held-out pair, the model produced 9 clusters with a mean absolute error of €313 (about 3.1% of mean cost), a bias of only +€12 (0.1%), and an R^2 of 0.31.

The pipeline was then extended to the full qualifying dataset. An empirical diagnostic showed that O-D pairs with 30-50 historical flights showed genuine route variation only 18% of the time, well below the roughly 20% threshold set for treating a flight-count band as reliable, so the minimum flight-count entry threshold was raised from 30 to 50. At this threshold, 4,558 O-D pairs qualified (536,520 flights), of which 2,494 (54.7%) showed genuine route or cost alternatives.

Two bugs were found and fixed during development: a column filter that leaked latitude, longitude and altitude data into the clustering distance metric, and a cluster-size guard that was not being enforced before the

clustering step ran. Fixing both changed the reported cluster counts and simultaneously improved validation accuracy, evidence that the fixes corrected genuine errors rather than just changing bookkeeping.

The result is a validated, generalisable simulator function that returns ranked route alternatives with predicted cost, duration and error bounds for over half of the O-D pairs with sufficient historical data.

2. Background and Aims

Background. Mercury, an air transport mobility simulator developed at the University of Westminster (Delgado et al., 2026), simulates European air transport mobility and already supports multiple flight plans per O-D pair, rather than assuming a single planned route. The practical limitation is that the flight-plan alternatives available to it are derived from historical data several years old, and some newer O-D pairs have no alternatives at all, since none have yet been derived from more recent data. Airlines and air traffic control regularly assign different routes for the same city pair, driven by weather, traffic flow, and North Atlantic Track allocation on long-haul routes, so this variation needs to be re-derived as new historical data becomes available. This project set out to build a pipeline that quantifies that variation directly from historical flight-plan data, and to establish the method needed to keep it current.

Before starting, the working assumption was that clustering flights by geographic waypoints would be enough to separate genuine route alternatives. In practice, flight information region (FIR) crossings, combined with the distance flown within each FIR, proved a more robust signal: it captures both which corridor a flight took and how that corridor was flown, without needing raw trajectory coordinates.

Aim. To characterise the real-world space of route and cost variation per O-D pair from historical EUROCONTROL flight-plan data, as a basis for more realistic route-alternative modelling in Mercury-style simulations.

Objectives:

1. Join FIR-distance and flight-plan data on flight ID.

2. Build a three-layer clustering pipeline: O-D pair, then FIR-crossing signature, then within-signature distance.
3. Attach a deterministic cost model, built from published ATC and fuel rates, to each cluster after clustering; cost is not itself a clustering feature, only FIR signature and distance are.
4. Validate the pipeline against a held-out O-D pair not used in development.
5. Extend the validated pipeline to the full dataset, choosing the entry threshold from evidence rather than a default.

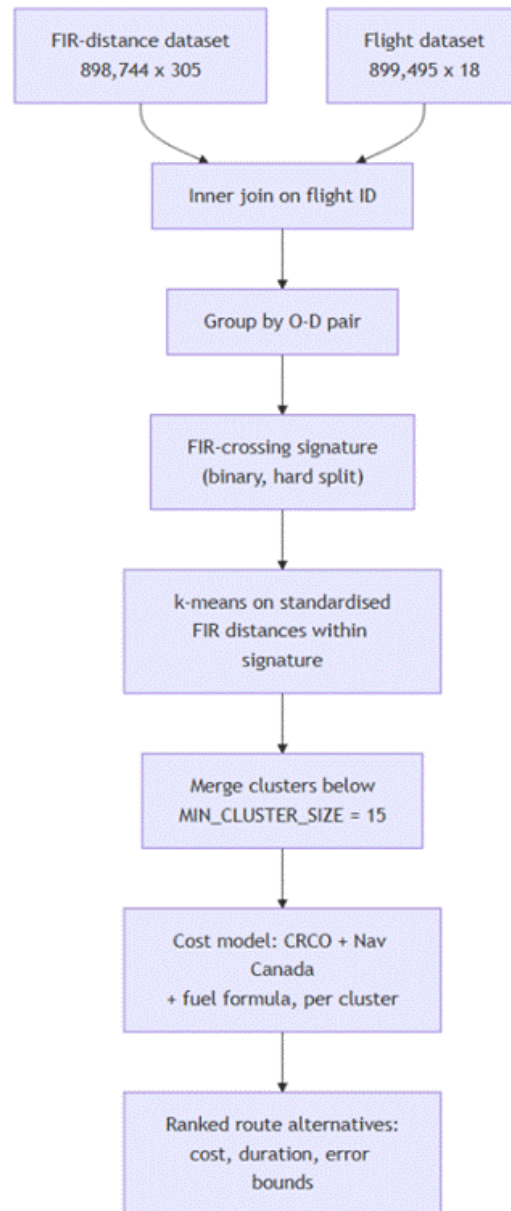
Impact on my own learning: The clearest lesson was how easily a pipeline that runs and returns plausible numbers can still be wrong underneath. Both bugs fixed during this project, the FIR-column filter and the unenforced cluster-size guard, produced output that looked reasonable on its own. Only comparing cluster counts and validation metrics before and after the fix showed the earlier numbers were not just imprecise, but actually incorrect. This changed how I check my own code. I now treat "it runs and the output looks sensible" as a weak signal, and try to verify each filter and threshold against what it is actually meant to do, rather than against whether the final numbers seem plausible.

A second, more technical lesson came from the clustering design itself. Splitting FIR-signature grouping (a hard, deterministic split) from distance-based k-means (a soft, data driven split) worked far better than a single flat clustering pass would have, because the two stages capture genuinely different kinds of variation: which corridor a flight took versus how it flew within that corridor. Collapsing them into one distance metric would have blurred a categorical difference and a continuous one together. It taught me to think about what kind of variation a feature actually represents before deciding how to cluster on it, rather than feeding everything into one distance function.

3. Methods

Two EUROCONTROL datasets covering September 2023 were used: a flight information region (FIR) distance dataset (898,744 rows, 305 columns, one column per FIR/UIR giving distance flown in nautical miles) and a flight dataset (899,495 rows, 18 columns, including origin, destination, aircraft type and duration). The two were

inner-joined on flight ID, matching 898,744 flights (0.08% dropped). No human participants were involved; the data is historical, aggregated air traffic control flight-plan data with no personal information, so no ethical review was required.



A three-layer k-means approach was used for clustering. Flights were first grouped by O-D pair. Within each pair, a first k-means pass (k chosen by silhouette score, range 2-5) split flights by their binary FIR-crossing signature. A second k-means pass then split each signature group further by standardised FIR distances, capturing within-corridor variation. Any resulting cluster smaller than 15 flights was force-merged into its nearest neighbour by centroid distance, to prevent fragile clusters driven by a handful of flights. A pair is

counted as showing genuine route variation when this process assigns it more than one cluster ($k \geq 2$); a pair that resolves to a single cluster is treated as having one dominant route with no meaningful alternative. This is the definition behind the "% of pairs with genuine alternatives" figures reported in Results.

Each cluster's predicted cost combines three components. En-route ATC charges follow the EUROCONTROL/CRCO formula, $(\text{distance_km} / 100) \times \sqrt{\text{MTOW_tonnes} / 50} \times \text{unit_rate}$, summed per FIR/UIR crossed using each state's published unit rate (FIR and UIR pairs from the same air navigation service provider are consolidated into one billing zone before charging). Nav Canada charges use a separate distance-based formula for domestic sectors, with Gander Oceanic charged as a flat fee instead, per Nav Canada rules. Fuel cost is cruise fuel flow (kg/h, by aircraft type) multiplied by flight duration and jet fuel price (€0.81/kg, EIA US Gulf Coast FOB, September 2023). MTOW and fuel-flow figures come from hardcoded lookup tables for 25 aircraft types (ICAO FEAT/EcoScope data), since OpenAP, the library that would normally supply this, is not usable due to a NumPy 2.0 incompatibility on Google Colab.

The dataset also contains both planned and realised trajectories for the same flights: the FIR-distance dataset above reflects the filed flight plan, while separate actual/filed trajectory files (actual FIR crossings and lat/lon points, plus the filed lat/lon points) record what was actually flown. Two comparisons were built from this. The first compares each flight's realised cost, duration and distance against the mean of its own cluster's planned values, the cluster representative: this is the noise the simulator needs to add around that single stored representative, since only one representative trajectory is kept per cluster, not one per flight. It is aggregated per cluster (mean, standard deviation, 5th/95th percentile) so it can be sampled from directly as simulator input. The second compares each flight's realised trajectory against its own filed plan, independent of which cluster it falls into. This measures how reliably route planning itself predicts what is actually flown, and can be compared across clusters and O-D pairs to see where planning is less certain.

Both comparisons use the same delta calculation, just against a different planned reference. For each flight, per metric (distance, duration, fuel):

- Comparison 1 (vs cluster representative): $\text{delta} = \text{actual_value} - \text{mean}(\text{planned_value for all flights in that cluster and aircraft type})$

- Comparison 2 (self-comparison): $\text{delta} = \text{actual_value} - \text{that same flight's own planned_value}$

Both are then aggregated by group (cluster and aircraft type for comparison 1; cluster, or O-D pair pooled across clusters, for comparison 2) into mean, standard deviation, and 5th/95th percentile, which is what the simulator or a robustness analysis would sample from or read off directly. Both comparisons have been run to completion on the three training pairs; comparison 2 has additionally been run across the full qualifying dataset, both pooled and split by carrier type (Results, Discussion).

The dataset's own carrier-type field (STATFOR Market Segment) was found to be unpopulated for every flight in the September 2023 export, so carrier type (low-cost vs full-service) was instead derived from the AC Operator field, a three-letter operator code. A hand-curated lookup table maps the most common operator codes (identified from the full dataset's actual operator frequency counts) to low-cost or full-service; operators not in the table, including the single largest code (an apparent placeholder for unidentified or general-aviation operators, not a real airline), are left unclassified rather than guessed. This is a best-effort classification, not sourced from an official registry, and leaves a substantial share of flights unclassified (Results).

The pipeline was validated on EGLL-LGAV, an O-D pair held out from all development, comparing predicted against actual cost. Analysis was carried out in Python (pandas, scikit-learn) in Google Colab for the main pipeline, with a local Python script (Plotly, SciPy) used for route-map visualisation.

4. Results

Training sample

O-D pair	Flights	Clusters
EGLL-KJFK	659	16
LEBL-LEPA	685	1
LPPT-EDDB	90	1

The transatlantic pair (EGLL-KJFK) showed the most variation, consistent with North Atlantic Track assignment changing day to day. The two short/medium-haul pairs showed no meaningful route variation, as expected for routes with limited airspace options.

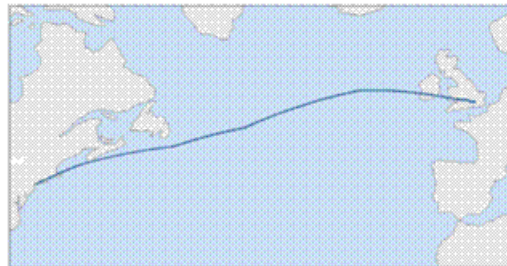
Worked example (EGLL-KJFK, cluster 9)

To illustrate what one cluster actually represents: cluster 9 is EGLL-KJFK's largest (123 of 659 flights). Its most common aircraft type is the B777-300ER (B77W, 41 flights); values below are for that type within the cluster.

Metric	Value
Cluster	C9, representative aircraft type B77W

Metric	Value
FIRs crossed	EGTT FIR/UIR (UK, ~135-266 nm), EGGX (Shanwick Oceanic, ~1,048 nm), CZQM (Moncton, ~1,257 nm), CZQX (Gander Oceanic, ~1,626 nm, flat-fee zone under Nav Canada rules)
n flights in cluster	119 total (41 of type B77W)
Mean distance	3,030 nm (B77W); 3,032 nm pooled across all aircraft types in the cluster
Mean duration	6.94 h (B77W); 6.98 h pooled
Mean fuel	54,147 kg (B77W)
Mean cost	€52,026 (B77W)

EGLL-KJFK | Cluster 9 representative actual trajectory
 ECTRL 265340206 (B77W) — Δdist +25 nm | Δdur -0.007 h



The flight shown (ECTRL 265340206, B77W) is the representative real flight nearest this cluster's centroid. Its actual trajectory differed from the centroid by only +25 nm and -0.007 h, showing how tight the within-cluster noise is for this particular cluster (the first comparison described above).

Note: these figures are from a fresh rerun with the corrected ATC rate table (Discussion), confirmed via `flight_routes` after the library conversion. The cost figure is unchanged from the pre-correction value

(€52,026): this cluster only crosses UK and Canadian airspace (EGTT/EGGX/CZQM/CZQX), none of which were among the 10 corrected FIRs, so it was never actually affected by the rate error, only routes crossing continental Europe were. The small changes in pooled n/distance/duration (119 vs the earlier 123, 3,032 vs 3,033 nm) reflect ordinary clustering variation between independent runs, not the rate fix.

Held-out validation (EGLL-LGAV, n=225)

Metric	Value
Clusters	9
MAE	€313 (3.1% of mean cost)
Bias	+€12 (0.1%)
R ²	0.31

Entry threshold selection

Flights per pair	% of pairs with genuine alternatives
30-50	18.0%
50-75	46.5%
75-100	55.9%
100-150	59.3%
150-200	66.3%
200-300	67.2%
300-500	58.2%
500+	39.0%

The 30-50 flight band showed real route variation in only 18% of pairs, below the 20% threshold set in advance for treating a band as reliable, so the minimum flight-count threshold was raised from 30 to 50 flights per pair.

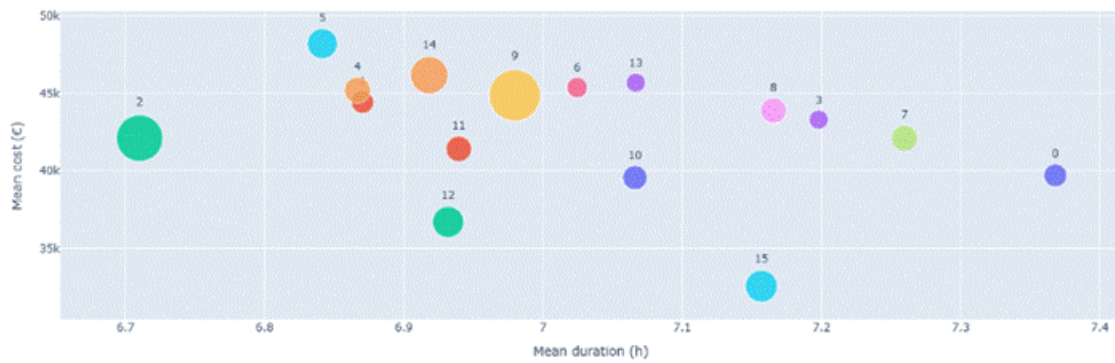
Full-dataset extension

At the revised threshold, 4,558 O-D pairs qualified, covering 536,520 flights. Of these, 2,494 pairs (54.7%) showed genuine route or cost alternatives, up from 40.8% at the original threshold of 30, confirming that the pairs excluded by the higher threshold were disproportionately single-route pairs rather than genuine alternatives being discarded. Aircraft type and cost data were available for 97.2% of flights in the full dataset.

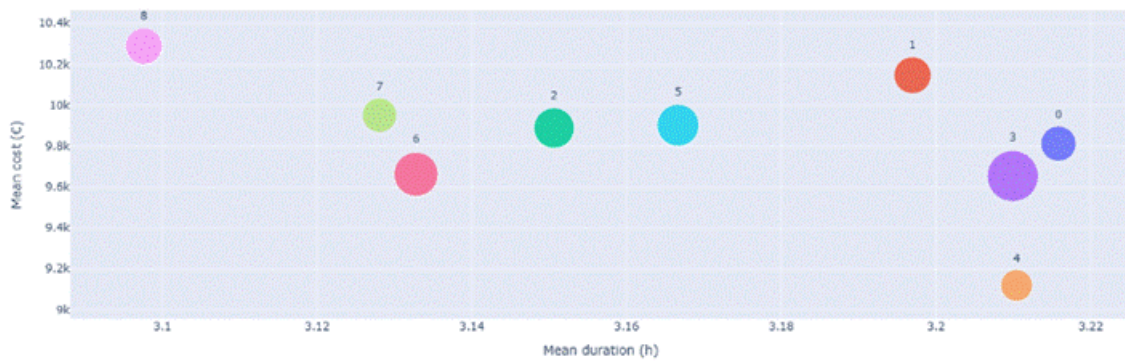
Supporting visualisations produced alongside these results include a world map of route-alternative density by O-D pair, FIR-usage heatmaps per pair, and cost-versus-duration scatter plots per cluster.



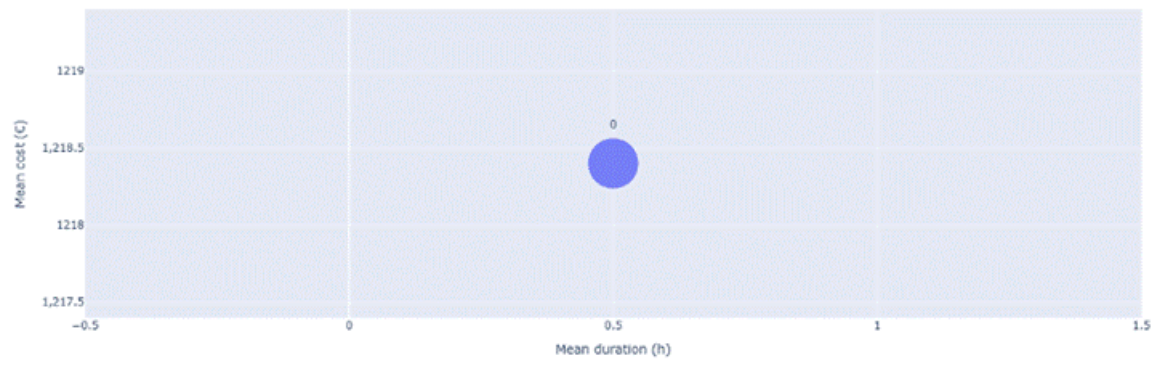
EGLL-KJFK: route alternatives



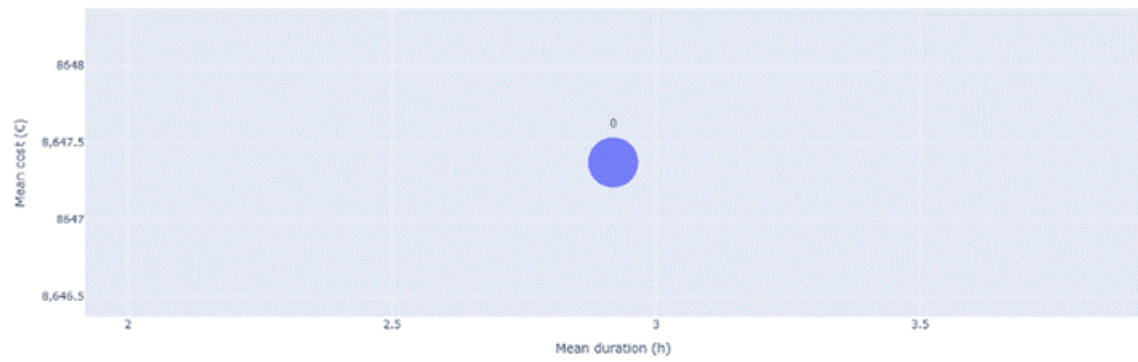
EGLL-LGAV: route alternatives - EGLL-LGAV validation



LEBL-LEPA: route alternatives



LPPT-EDDB: route alternatives



Heatmap showing mean nm values for 16 cell lines (C0 to C15) across 8 conditions (C0P1 to K2P1). The color scale ranges from 0 (dark green) to 2500 (dark red). The cell lines and their sample sizes (n) are listed on the y-axis. The conditions are listed on the x-axis. The heatmap shows that cell lines C11, C12, and C13 generally have higher mean nm values (yellow/orange) compared to others, while C0, C1, and C2 have lower values (dark green).

Heatmap showing the mean number of reads (nm) for 16 different methods (C0 to C8) across 16 different datasets (EBR, EDU, EST, LAA, LGO, LPP, LOGG, LIB, LHM, LIA, LOU, LQB, LSS, LWS, LXA). The color scale ranges from 0 (dark green) to 700 (red).



Full-dataset planned-vs-realised variance and carrier-type comparison

The per-flight self-comparison analysis (Methods, Discussion) was extended from the three training pairs to all qualifying O-D pairs with complete actual and filed trajectory data: 4,531 of the 4,558 qualifying pairs (536,517 of 536,520 full-dataset flights matched to both records). Weighted by pair size, flights across the whole dataset flew on average 7.3 nm less than filed, with essentially no net duration bias, and 25.8 kg more fuel than planned. This average distance bias is consistent in direction and order of magnitude with two of the three training pairs (LEBL-LEPA: -12.5 nm; LPPT-EDDB: -2.9 nm). This suggests the training sample was not unrepresentative of the wider dataset.

Individual pairs varied considerably around this average. The largest single-pair mean biases were LTFJ-ORER (+111.3 nm, n=71) and UBBB-UUEE (-107.5 nm, n=90); the widest spreads (standard deviation of the distance delta) were LLBG-LPPT (360.7 nm, n=53) and OJAI-EGLL (344.7 nm, n=74). Three of the five widest-spread pairs shared LTFM (Istanbul Airport) as their departure airport, initially suggesting route-planning from this airport is unusually variable. Investigating the individual flights behind this showed otherwise: in each of the three LTFM pairs, the single most extreme flight had an actual trajectory that landed exactly at the correct destination, but a filed trajectory record that matched a completely different, much longer flight (to Jakarta, Boston, and Douala respectively, rather than Muscat, Brussels, and Malta), a data-linkage error in the raw filed-trajectory export for that specific flight ID, not a real difference in flight planning. Excluding just that one mismatched flight per pair reduced the standard deviation by 6-17x (LTFM-OOMS: 336.2 to 25.6 nm;

LTFM-EBBR: 265.9 to 15.6 nm; LTFM-LMML: 241.5 to 38.9 nm), collapsing all three to an unremarkable spread in line with other pairs in the dataset. The apparent Istanbul-specific pattern was therefore a data artefact affecting a handful of individual flight records, not a genuine finding about that airport's route planning.

Splitting the self-comparison by carrier type (Methods) reached 62.9% of flights (41.3% full-service, 21.5% low-cost), with the remaining 37.1% unclassifiable from the AC Operator field alone (Methods). Low-cost carriers showed a tighter spread than full-service on both distance (std 25.4 nm vs 29.9 nm) and fuel (std 351 kg vs 556 kg), despite a similar mean distance bias (-7.9 nm vs -6.9 nm). Full-service carriers also showed a larger mean fuel overrun (+49.7 kg vs +1.1 kg for low-cost). This points to low-cost carriers' flight plans being modestly more predictable in practice, though the gap is one of degree rather than a clear qualitative difference, and the sizeable unclassified share limits how far the comparison can be pushed.

5. Discussion

The pipeline achieved its aim of generating a realistic, generalisable set of route alternatives from historical data. A Random Forest model was also tested as an alternative to the formula-based cost model: it achieved a near-perfect in-sample R^2 of 0.9996, but failed on the held-out O-D pair because it had effectively memorised which FIRs mattered for the three training routes and could not generalise to FIRs it had not seen. The formula-based model, built from published ATC and fuel rates rather than learned per-route, generalised correctly instead. This comparison is itself a useful finding: for a simulator meant to work on unseen routes, a method that generalises by construction is more valuable than one that fits the training data well but cannot extrapolate.

The three-layer clustering design worked as intended, separating corridor-level route choice (which FIRs a flight crossed) from within-corridor variation (how far it flew within each FIR, for example due to different North Atlantic Track assignment). This distinction would have been lost using a single clustering pass. Beyond feeding Mercury directly, the same cluster structure supports comparative analysis across O-D pairs, for example ranking pairs by how much route or cost variation they exhibit, or identifying which pairs are effectively single-route versus genuinely multi-route, without needing bespoke analysis per pair. The comparison between

each flight's realised cost, duration and distance and its own cluster's planned representative found that the residual is not random noise. Where two clusters share an identical FIR signature, their real difference concentrates in the oceanic-crossing FIRs charged by distance (Moncton, Montreal, Shannon), differing by up to several hundred nautical miles, consistent with different North Atlantic Track assignments on different days. This is the noise distribution the simulator should sample from around each cluster's representative, not a modelling error to be reduced away.

EGLL-KIPK: noise (actual - reference) by cluster and AC type





The complementary comparison looks at each flight's realised trajectory against its own filed plan, rather than the cluster representative. It measures how reliably route planning itself predicts what is flown, independent of clustering, and that reliability can be compared across clusters and O-D pairs. Pooled across all clusters,

EGLL-KJFK flew on average 1.5 nm further than filed (std 6.6 nm, 5th-95th percentile -7.6 to +10.5 nm) and 235 kg more fuel than planned (std 480 kg). LPPT-EDDB showed a small negative bias instead (-2.9 nm on average, but with the widest spread of the three pairs, std 20.0 nm). LEBL-LEPA had the largest consistent bias, flying 12.5 nm less than filed on average, though with a tighter spread than LPPT-EDDB (std 10.7 nm). The short-haul pair is therefore the most predictable from its filed plan, and the medium-haul continental pair (LPPT-EDDB) the least, despite having far fewer FIR crossings than the transatlantic pair. This suggests route-planning uncertainty does not simply track route length or corridor count.

Extending the self-comparison to the full dataset (Results) confirms the training-sample pattern rather than overturning it: the -7.3 nm pooled distance bias sits within the range already seen for LEBL-LEPA and LPPT-EDDB. The Istanbul-origin pairs' apparently unusual spread (Results) turned out not to be a route-planning finding at all: a single flight per pair had a filed trajectory record belonging to an entirely different, much longer flight, and excluding that one flight per pair collapsed the spread to an ordinary level. This is a second, independent example of the same lesson as the FIR-column filter bug below: a summary statistic that looks like a real, reportable pattern can instead be a single corrupted input dragging a std-based metric far from what the underlying data actually shows, and is only caught by inspecting the individual records behind it rather than trusting the aggregate number. As suggested by the academic partner, splitting by carrier type found low-cost carriers' flight plans to be modestly more predictable than full-service carriers' on both distance and fuel, despite a similar average bias. This is a small but genuine difference, limited by the 37.1% of flights that could not be classified from the available operator data.

Two bugs were found and fixed during the project. First, the filter used to select FIR columns matched on data type (float64) rather than column name, which meant latitude, longitude and requested flight-level columns were incorrectly included in the clustering distance calculation. Second, a minimum cluster size of 15 flights was not being enforced on clusters formed before the k-means step ran, allowing clusters of one to four flights to survive. Fixing both bugs changed the reported cluster counts and simultaneously improved held-out MAE, bias and R^2 together. Because all three metrics improved at once rather than trading off against each other, this is treated as evidence that the fixes corrected genuine errors in the pipeline rather than just changing which flights were grouped together.

The entry threshold for the full-dataset extension (minimum 30 vs 50 flights per O-D pair) was set using diagnostic evidence rather than an arbitrary cutoff: pairs with 30-50 historical flights showed genuine route

variation only 18% of the time, compared with 46-67% for pairs with more flights. Interestingly, the very highest-traffic pairs (500+ flights) showed lower variation again (39%), plausibly because very high-frequency pairs tend to be shorter, single-corridor routes with less opportunity for route variation than busy long-haul corridors.

There are some limitations to note. The pipeline was validated against only one held-out O-D pair, so the reported MAE, bias and R^2 should be read as indicative rather than definitive; a larger held-out set would strengthen this claim. A small number of flight IDs in the raw filed-trajectory export have a filed record that does not correspond to their actual O-D pair (Results); these were identified individually where they distorted a pair's self-comparison spread, but no systematic search for further instances of this issue has been carried out, so isolated cases may remain elsewhere in the full dataset. A full audit of the ATC rate table against EUROCONTROL's September 2023 monthly adjusted unit rate table found that 10 of the 18 rate entries (not just the two originally flagged as estimated) were incorrect, by margins ranging from 6% to over 100%; these have since been corrected and the full pipeline rerun end to end. This confirms the same 4,558 qualifying pairs and 54.7% showing genuine alternatives as before the fix. This is expected: clustering runs on FIR distance and signature only, not cost, so the correction changes absolute ϵ cost accuracy without affecting which pairs show alternatives. The silhouette score for EGLL-KJFK is low (0.13) because several of its 16 clusters share an identical modal FIR signature and are distinguished only by within-corridor distance; this reflects a limitation of using a binary-signature distance metric to judge cluster quality, not an error in the clusters themselves, which were confirmed as genuinely distinct by inspecting mean per-FIR distances directly.

These findings are broadly consistent with the wider literature on flight-path clustering: Song et al. (2020) also use partition-based clustering methods on route-signature and waypoint features to separate genuine route corridors from noise. This supports the choice of a signature-based approach over raw trajectory coordinates here.

6. Conclusion and Recommendations

This project shows that a substantial share of O-D pairs, 54.7% of those with at least 50 historical flights, exhibit genuine route or cost variation that a single-planned-route assumption would miss. This is a meaningful basis for adding realistic route-alternative distributions to Mercury-style simulations, rather than assuming a single fixed route per O-D pair.

The key recommendation from this work is methodological: a minimum flight-count threshold of 50 per O-D pair should be used before treating an O-D pair as a candidate for route-alternative modelling, based on diagnostic evidence rather than a round-number default. Below this threshold, an O-D pair should be treated as single-route by default, since apparent variation below 50 flights is more likely to be noise than signal.

Future work:

- Systematically screen the full dataset for other flights with a mismatched filed-trajectory record (Discussion), rather than relying on ad hoc discovery via per-pair outlier inspection, since the three found here were only surfaced because they happened to distort a pair's reported spread enough to stand out.
- Migrate from ad hoc file-based caching (CSV/parquet on Google Drive) to a proper database as the dataset scales beyond a single month.
- Repeat the analysis on other months to check whether the clustering results and thresholds are stable across seasons, or specific to September 2023.

7. Dissemination

The main novel technique in this project is the three-layer clustering design itself, splitting route characterisation into corridor choice (FIR signature) and within-corridor variation (distance), which allowed genuinely different routing behaviour to be distinguished from noise. A second technique worth noting is the debugging method used to validate clusters that appeared to be duplicates: rather than assuming clusters sharing a modal FIR

signature were redundant, their mean per-FIR distances were compared directly, which showed they differed by hundreds of nautical miles on oceanic-crossing FIRs specifically, confirming they represented real route variants rather than clustering noise.

The clearest lesson from this project is that a filter or threshold bug can hide inside a pipeline that still runs and produces plausible output. The FIR-column filter bug did not cause an error or an obviously wrong result; it silently corrupted the clustering distance metric while the pipeline kept running and producing numbers that looked reasonable. The lesson is that filters and thresholds need to be checked end to end against their intended logic, not just spot-checked against whether the final output looks sensible. The same lesson reappeared late in the project in a different form: the unusually wide self-comparison spread for Istanbul-origin pairs looked like a genuine, reportable finding until the individual flights behind it were inspected directly, at which point it turned out to be one mismatched data record per pair rather than a real pattern. A summary statistic that looks interesting is worth checking against the raw records before it is trusted, not just against whether it seems plausible.

A second lesson concerns threshold decisions specifically: the minimum flight-count threshold for the full-dataset extension was set using a diagnostic (percentage of pairs showing genuine variation, by flight-count band) rather than choosing a round number in advance. This made the final choice defensible with evidence rather than a guess.

Writing the documentation first turned out to clarify my own understanding of the pipeline before I restructured it into modules, which probably made the eventual library conversion cleaner. It taught me that a supervisor's instruction can carry practical value that isn't obvious until you're partway through it.

This report is the primary dissemination output of the project. The full codebase, packaged as an installable Python library with tests and documentation, is publicly available under a GPL-3.0 licence at <https://github.com/3nd03/flight-route-characterisation>. I would welcome the opportunity to present this project further, should it arise

.

8. Research Team Reflection

This was a solo project, so the reflection below concerns the individual working relationship between myself and my supervisor rather than a peer team dynamic.

What went well: Luis provided clear guidance at key decision points while giving me autonomy to work and learn independently.

Problems: The size of the dataset was difficult to work with in practice, as it could not be shared directly with my supervisor. All processing therefore had to be carried out using cloud-based tools.

What enabled success or caused problems: Luis's subject knowledge and willingness to help made the project run smoothly. Frequent meetings and check-ins, scheduled around my study commitments, meant I did not fall behind.

What I would do differently: I would validate that filters and size thresholds are enforced end to end earlier in the pipeline, rather than discovering after several sessions that a bug had been silently affecting cluster counts. More generally, I would treat every filter and threshold as something to verify against its intended logic directly, rather than trusting that plausible-looking output means the logic is correct.

References

Delgado, L., Gurtner, G., Weiszer, M., Bolic, T., & Cook, A. (2026). Mercury: An open-source simulator for the evaluation of air transport mobility. *European Journal of Transport and Infrastructure Research*, 26(1).

Song, Y., Yu, K., & Young, S. (2020). Categorizing flight paths using data visualization and clustering methodologies. *Proceedings of the 9th International Conference for Research on Air Transportation (ICRAT 2020)*.