Exploring Trip Circuity in Columbia, SC

Xiwen Hao (STAT 531, University of South Carolina)

1. Why I cared about "circuity" in the first place

When I first moved from China to USA I quickly learned that Google Maps' "5 km" doesn't always feel like 5 km. Some drives loop around rail yards, creeks, or old mills; others shoot straight through on a freshly widened arterial.
That difference between the straight-line distance and the real path a vehicle must follow is captured by one deceptively simple ratio:

circuity = network distance ÷ Euclidean distance

Australian transport students call it *the squiggliness factor*. American planning professors (Levinson & El-Geneidy) gave it a fancier label, *the minimum circuity frontier*, and showed that commuters tend to organise their home–work pair so that this ratio is as low as possible .

Because our class got a rich SUMO simulation output—8 k healthcare-oriented trips with timestamps, lat/lon, speeds, waiting times, and even individual lane IDs—I thought it'd be cool to see:

- Is the Columbia street network as "efficient" (in circuity terms) as the Twin Cities or Portland?
- Can a mid-sized data set and a plain-vanilla machine-learning model predict that ratio to a useful precision?
- Does the literature's 1.2 'rule-of-thumb' still make sense in an era of sprawling suburbs and GPS routing?

2. Data wrangling

- Raw file: tripinfo_details.csv, 8 328 rows, one row per simulated trip from a random household cell to the closest medical facility.
- I kept only rows where all four coordinates were present and numeric. That wiped out 39 records—good riddance.
- A tiny UDF called haversine() gave me the airline distance in metres. Anything with EuclDist = 0 (a bug) or circuity > 5 (probably a teleporting ambulance) was kicked out.
- After cleaning I had 8 289 trips. Mean airline distance is 16.1 km; mean network distance is 21.9 km; so the naïve average circuity is 1.36—already lower than I expected.

I also engineered three extra columns:

Feature	Why I added it
AvgSpeed = RouteLength / Duration	Captures congestion and speed limits implicitly
One-hot dummies for facility type	To see if urgent-care centres in strip malls behave differently from big hospitals downtown
EuclDist_decile (0–9)	Quick way to stratify by trip length in later plots

3. Modelling without over-modelling

I didn't feel like going down a hyper-tuning rabbit hole. My learning curve experiments showed Random Forest generalises well after ~300 trees, so I froze those parameters and only scaled numeric features.

```
# scikit-learn pipeline in one breath
pipe = Pipeline([
    ('scale', StandardScaler()),
    ('rf',    RandomForestRegressor(n_estimators=300, random_state=42))
])
```

I still compared it against Gradient Boosting, SVM (rbf), AdaBoost, and a tiny two-layer stacking ensemble—mostly for curiosity.

Training–test split: 80 / 20, random seed 42.
Metrics: MAE, RMSE, R^2. Cross-validated RMSE (5-fold) for a sanity check.

4. What came out of the blender

Model	MAE	RMSE	R^2
Random Forest	0.010	0.037	0.969
Stacking (RF+GB → LR)	0.013	0.037	0.969
Gradient Boosting	0.024	0.043	0.957
SVM (rbf)	0.059	0.102	0.766
AdaBoost	0.201	0.232	–0.216

Numbers are unit-less (circuity is a pure ratio).

Take-aways

- A mean absolute error of ±0.01 means the forest knows circuity to within one percentage point(!).

- The scatterplot of predicted vs. actual ratios hugs the 45° line; residuals are basically Gaussian with a zero mean.
- Feature importances shout out the obvious heroes—RouteLength, EuclDist, and Duration. Facility type barely registers.

In practical words: if I know how long a trip *could* be and how long it *actually* takes, I already know 95 % of the story; land-use labels add pocket change.

5. How does Columbia compare to the literature?

- Columbia mean circuity = 1.36
  Levinson & El-Geneidy saw 1.18 for real commutes in Minneapolis and 1.19 in Portland.
- Why am I higher? Two guesses:
  1. Healthcare trips originate everywhere, not just from densely served suburbs. Many households in rural Richland or Lexington County only have one viable arterial that meets the Interstate in a long dog-leg.
  2. Street hierarchy differs. Minneapolis' historic grid extends deep into first-ring suburbs; Columbia has more curvy subdivisions with one or two exits.

Even so, my 1.36 is *way* below the 1.58 ratio that random short trips in the Twin Cities showed in the 2009 paper's Figure 3 . So, yes, people or planners (or SUMO's routing algorithm) still appear to "apply intelligence".

6. Visual nuggets

- Circuity vs. Route Length – a box-plot by decile drops from ~1.42 (shortest routes under 8 km) to ~1.28 (longest routes over 40 km). Same trend the 2009 paper hinted: short urban hops are forced to zig-zag around blocks; long freeway legs stay straight.
- Residual map (not shown here): the biggest under-predictions cluster near Lake Murray dam—turns out SUMO sometimes picks a lakeside detour that humans rarely drive.
- Learning curve – both training and CV RMSE flatten after ~3 000 samples, so the forest isn't data-starved but neither is it over-fit.

7. Policy reflections through a foreign student's lens

Back home in China, planners love mega-grid arterials: 100 m spacing, eight lanes each way—very direct but hostile to pedestrians. Columbia's network, in contrast, mixes 19th-century grids, mid-century cul-de-sacs, and modern arterials. The 1.36 average suggests that despite that patchwork, drivers don't suffer epic detours.

However, micro-level fixes could still shave seconds:

- Access management along suburban arterials—cutting extra U-turns can drop circuity by ~0.02 for local residents.
- Signal timing won't change the ratio but *does* affect perceived detour pain; a 50-second left-turn wait makes a "direct" route feel longer.
- More cross-parcel connectors in new subdivisions would fight the high circuity seen in the first 3 – 5 km distance bin.

8. Limitations I'm not hiding

1. SUMO isn't real life. The simulated driver may take the mathematically shortest path, not the psychologically preferred one.
2. No time-of-day tags. Morning congestion could push people onto indirect but faster freeways. I treat every trip as time-neutral.
3. Single-city focus. One metro area can't settle the frontier debate, though my numbers line up plausibly with the 22-city panel in the 2009 study.
4. No socio-economic variables. A high-income household may trade distance for school quality; my model is agnostic.

9. What I'd try next step

- Add OpenStreetMap link types (service road vs. trunk vs. residential) into the feature set.
- Run a mixed-effects model with random intercepts for origin census tracts—might capture local street morphology effects better than a global forest.
- Test against smartphone GPS traces if DOT ever releases anonymised pings; then I could compare "simulated ideal" routes to what flesh-and-blood drivers actually choose.

10. Final thoughts

Writing this report taught me that sometimes simple geometry plus a half-decent algorithm already tells a rich urban story. Circuity isn't just an engineer's curiosity; it hints at how forgiving—or punishing—a city's layout feels to its residents. And as an international student navigating between classes, Publix, and the dentist, I can confirm: less squiggle equals less stress.

References

Levinson, D., & El-Geneidy, A. (2009). *The minimum circuity frontier and the journey to work*. Regional Science & Urban Economics, 39, 732–738.