

When observer effort does not match expectations in opportunistic data

Annaëlle Bénard (LPO AuRA, LEHNA UMR 5023)

Thierry Lengagne (LEHNA UMR 5023)

Ecological context: wildlife-vehicle collisions

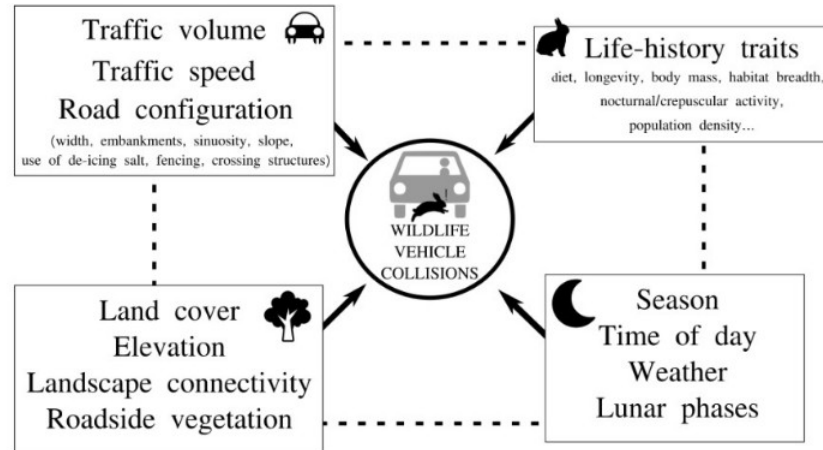


Grilo et al. (2020)

Birds : 200 millions / yr

Mammals : 30 millions / yr

Spatio-temporal patterns



Ecological context: wildlife-vehicle collisions

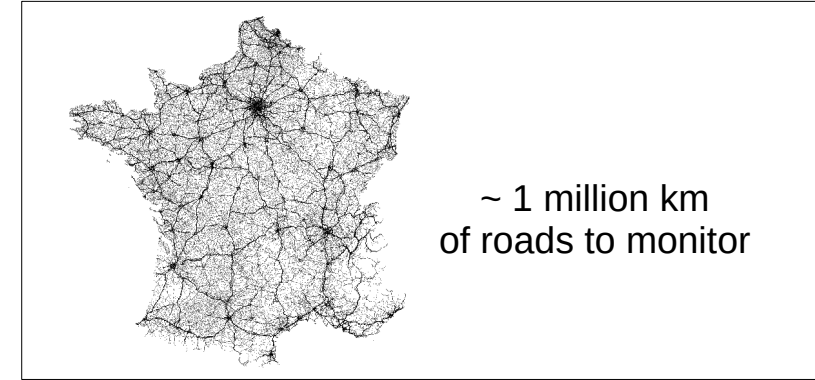
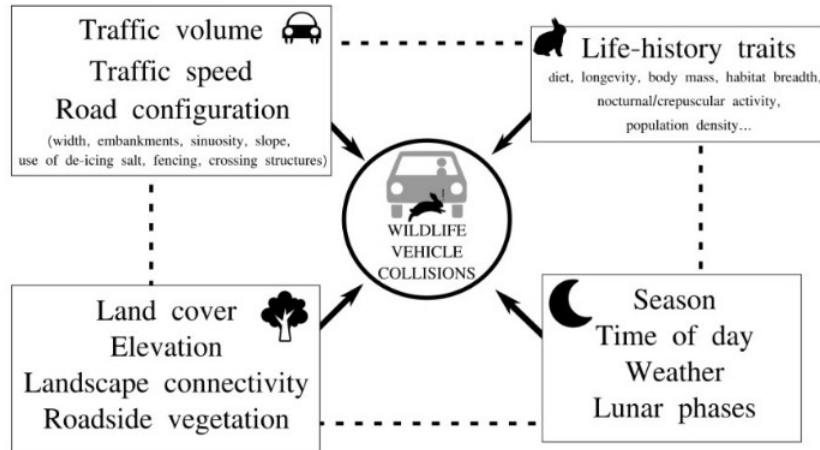


Grilo et al. (2020)

Birds : 200 millions / yr

Mammals : 30 millions / yr

Spatio-temporal patterns



Citizen science reports
→ opportunistic presence-only



Faune
France



The dataset: Faune-France *road mortality* in the Auvergne-Rhône-Alpes (AuRA) region

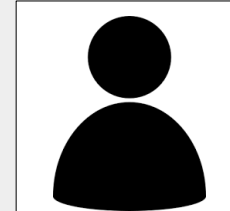


60 000 reports

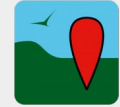
300 species
(mammal, birds, reptiles)

2680 participants

Profile of the typical FF participant (not roadkill-specific)



Male
40+ years old
Bachelor or above



Nature-oriented amateurs (farmer, bio teacher, vet,...)

Ecology professionals (consultant, NGO employee,...)

From internal LPO surveys and Charvolin, Joliveau & Pietrapoli (2025)

Average sociological profile in FF
≠
french population

The dataset: Faune-France *road mortality* in the Auvergne-Rhône-Alpes (AuRA) region

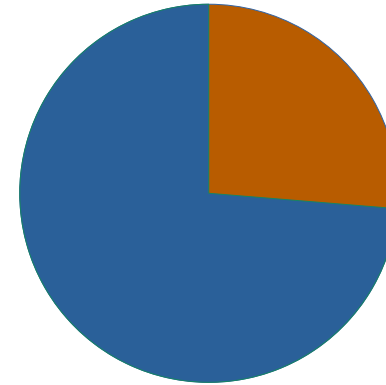


60 000 reports (2015 - 2024)

300 species
(mammal, birds, reptiles)

2680 participants

73 % of roadkill data
by 10 % of the participants



47% reported 1 or 2 collisions
median reports: 3



3.3% reported 100+ collisions

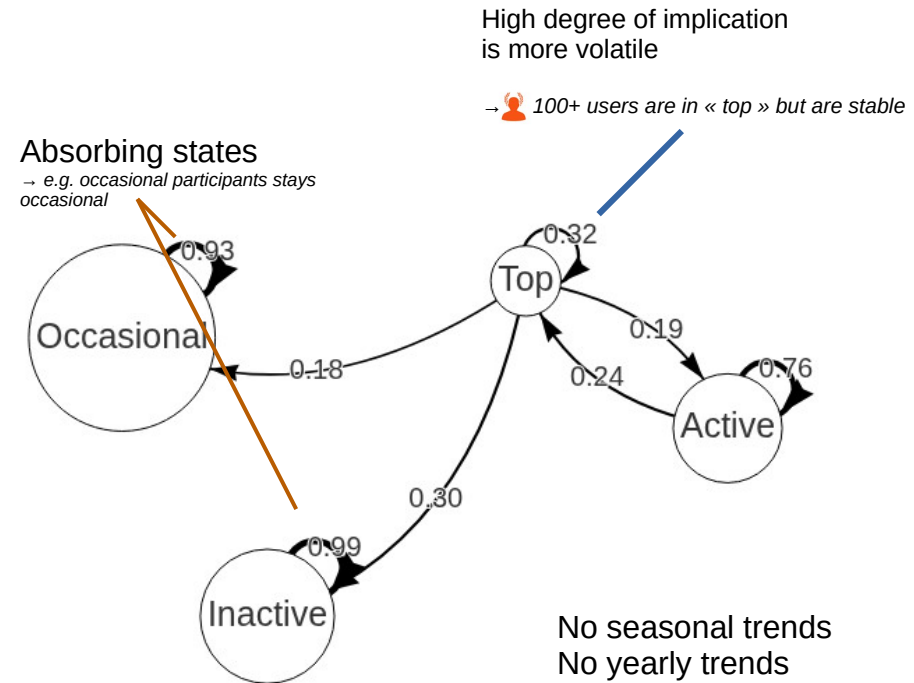
→ In AuRA, top ten participants = LPO employees

The dataset: Faune-France *road mortality* in the Auvergne-Rhône-Alpes (AuRA) region



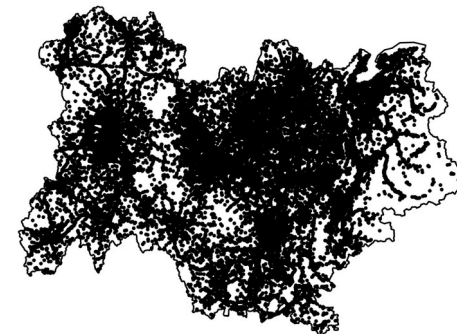
Hidden Markov Model of reporting states (France)

- number of reports per trimester
- « reporting state » is latent



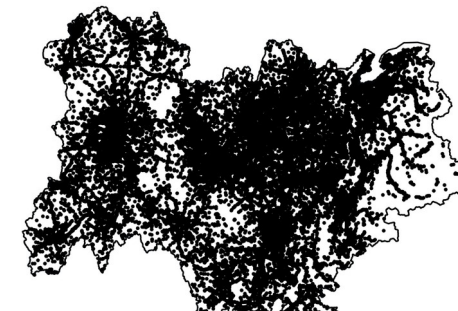
Research context

Infer sampling effort in opportunistic datasets with no absences



Research context

Infer sampling effort in opportunistic datasets with no absences

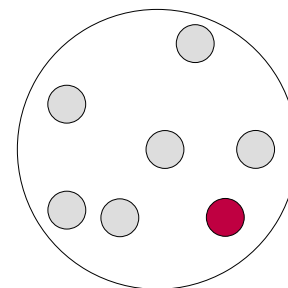


Independent proxy

- distance to roads
- trail/road density
- population density
- distance to settlements
- points of interest
- ...

- Try to find something independent of the ecological process
- Assumed to be a good proxy of actual latent sampling effort

Target-Group

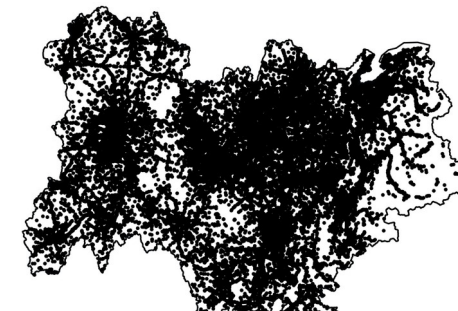


- All other species
- Other species of same taxa
- ...

- Even detection and reporting rates between species
- Combined occurrences of the target group species don't reflect a shared ecological niche, but where observers tend to go.

Research context

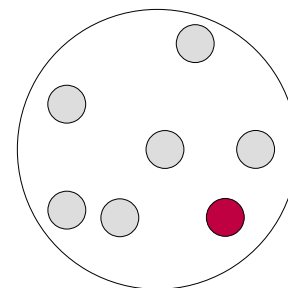
Infer sampling effort in opportunistic datasets with no absences



Independent proxy

- distance to roads
- trail/road density
- population density
- distance to settlements
- points of interest
- ...

Target-Group



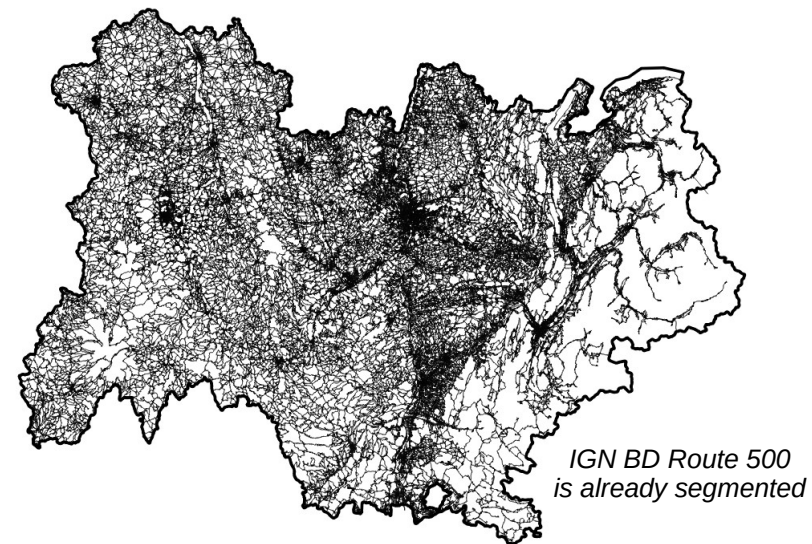
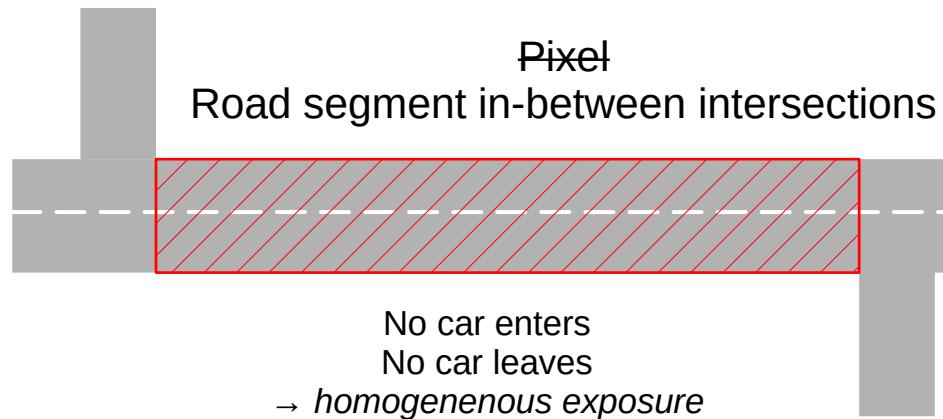
- All other species
- Other species of same taxa
- ...

Wildlife-Vehicle Collisions?

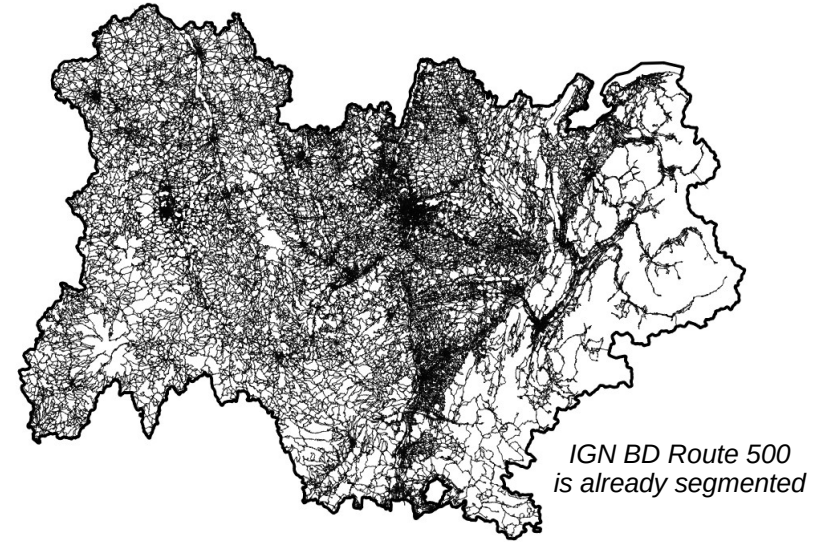
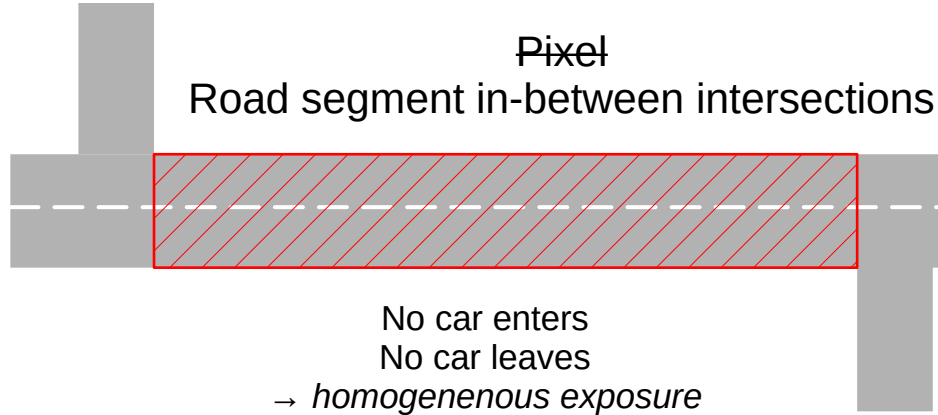
Observation process in FF: Visit of the road → Detection → Report

exposure

Spatial unit



Spatial unit



H1: Are road visits by participants explained by road traffic? (proxy approach)

Most intuitive proxy of number of visits / road of FF participants

H2: Can roadkill counts capture sampling effort? (TG approach)

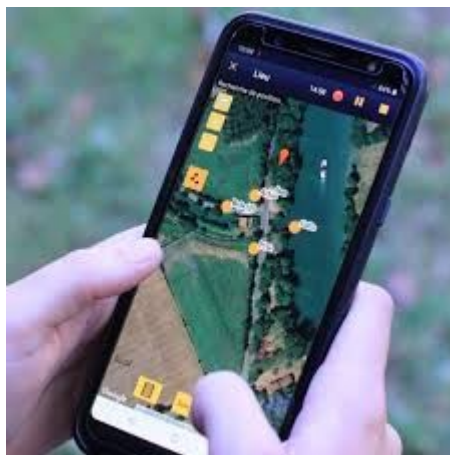
Test the hypotheses: true road exposure

30 drivers (FF users) over 2 months

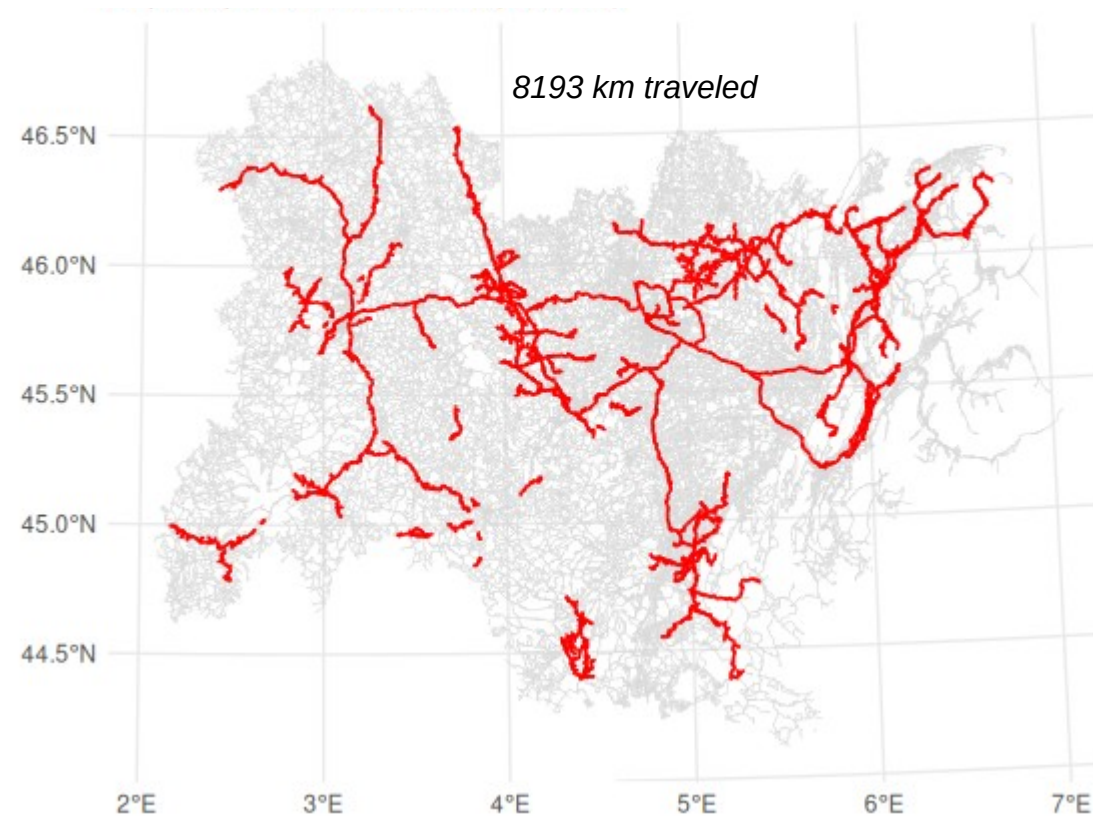
- including all  + handful of occasional participants

(LPO employees + active LPO members)

Record all car journeys
+ roadkill data



The *Transmorta* dataset



H1: Are road visits by participants explained by road traffic?



Annual Average Daily Traffic (AADT)
Known for <2% of roads

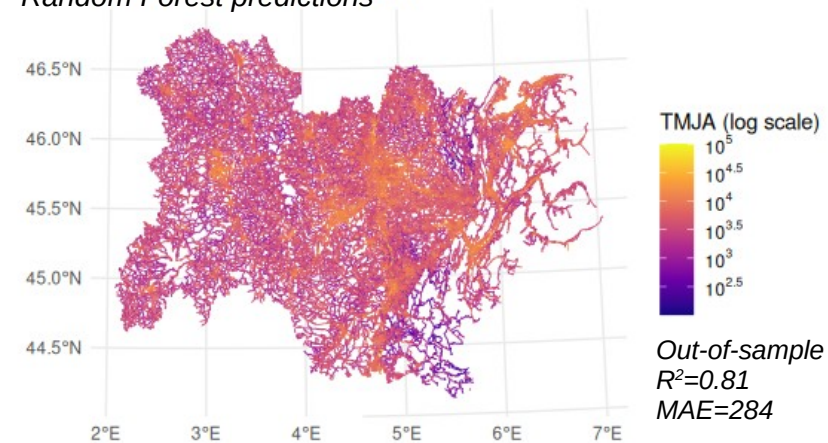
H1: Are road visits by participants explained by road traffic?



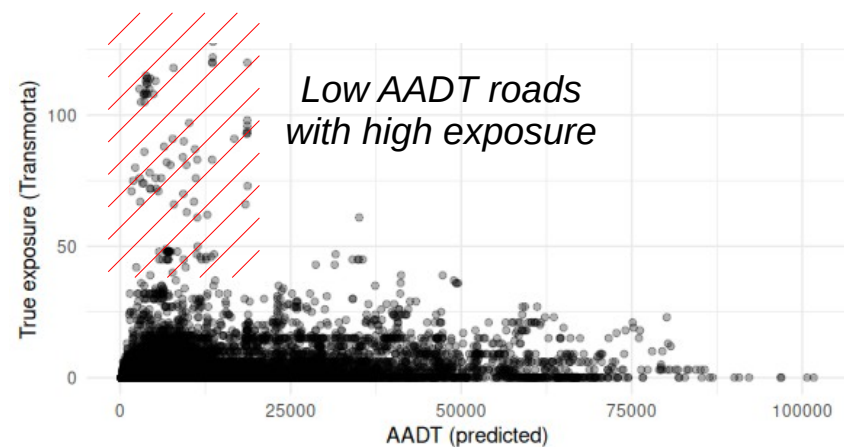
Annual Average Daily Traffic (AADT)
Known for <2% of roads

- Population density
- Road features (tolls, adm. class, number of lanes, ...)
- Average km traveled/car in France
- Graph embeddings

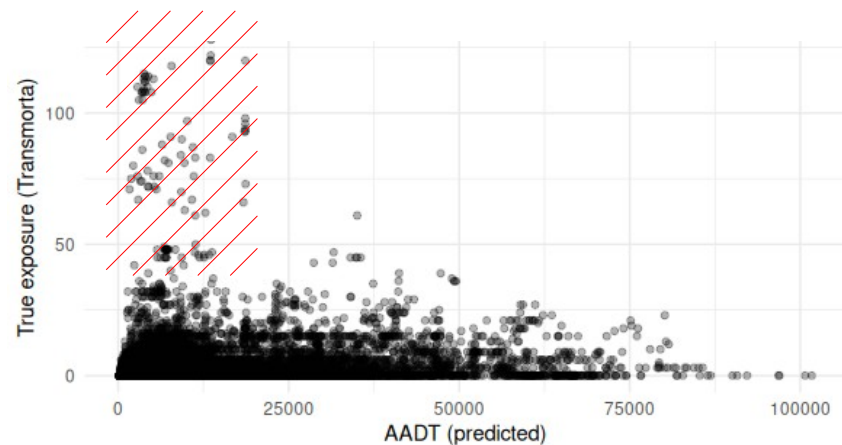
Random Forest predictions



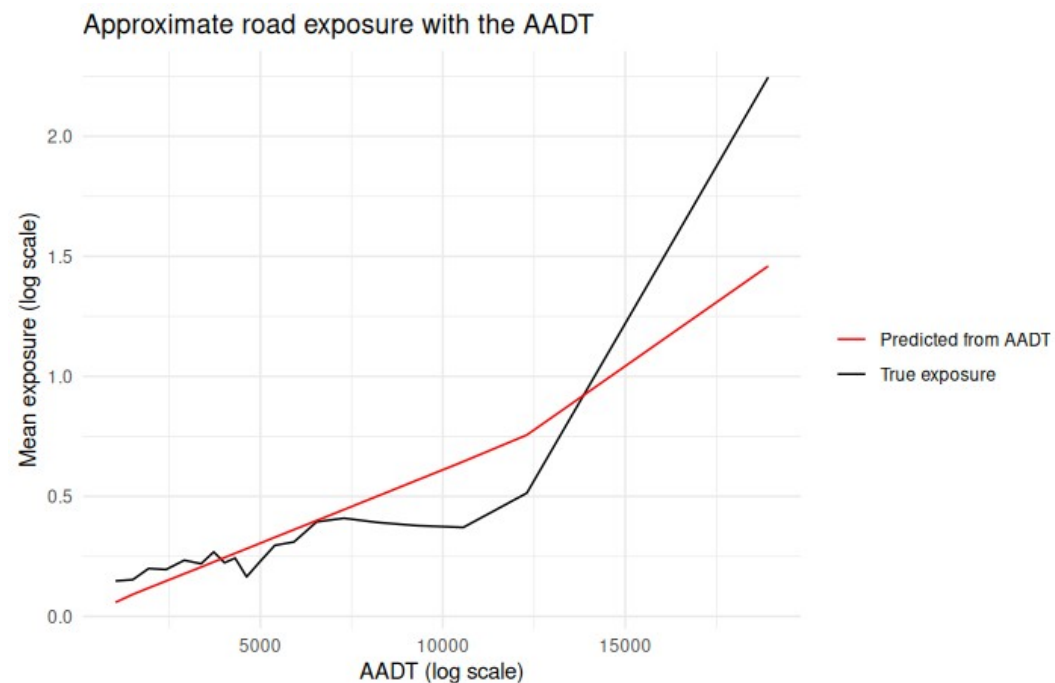
H1: Are road visits by participants explained by road traffic?



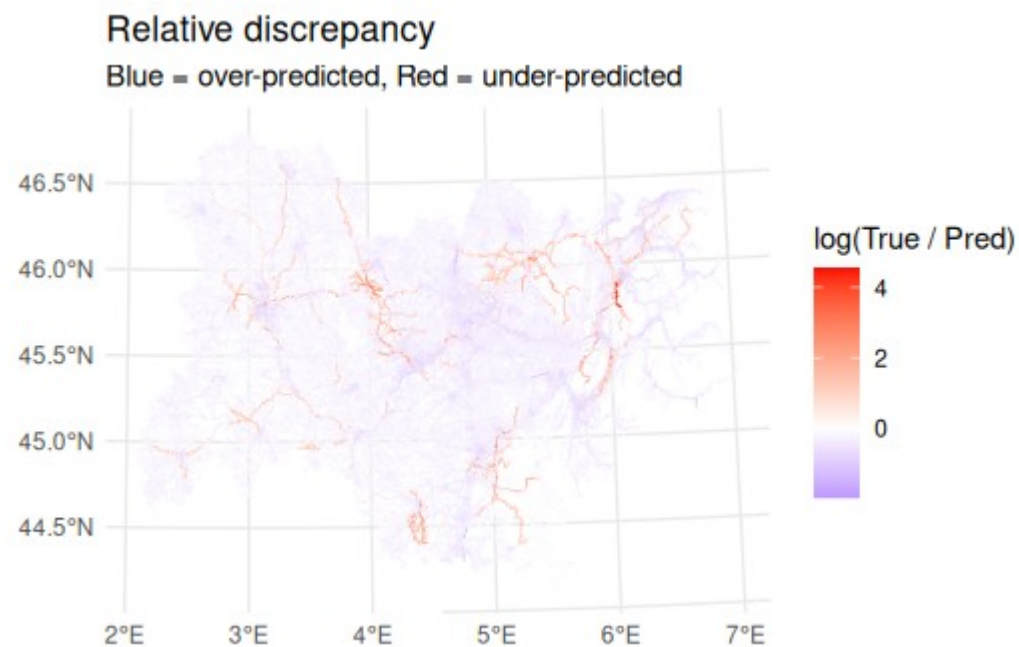
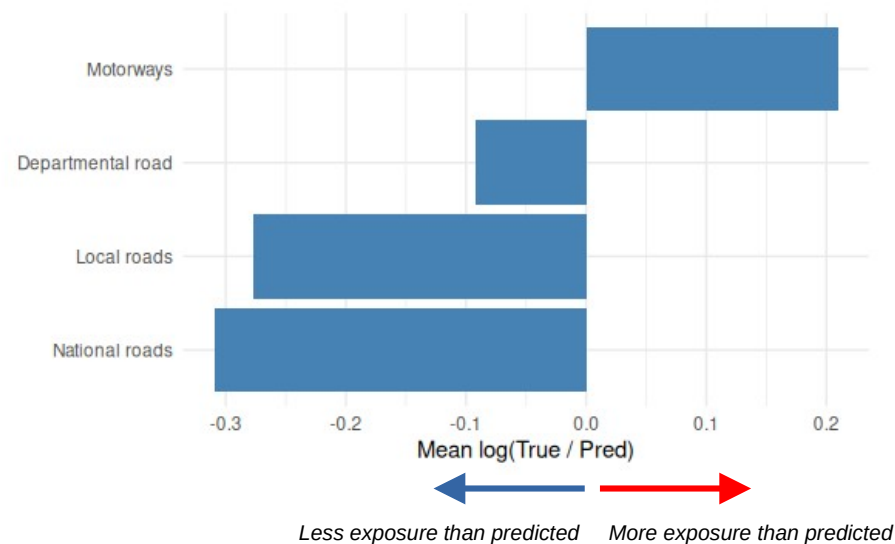
H1: Are road visits by participants explained by road traffic?



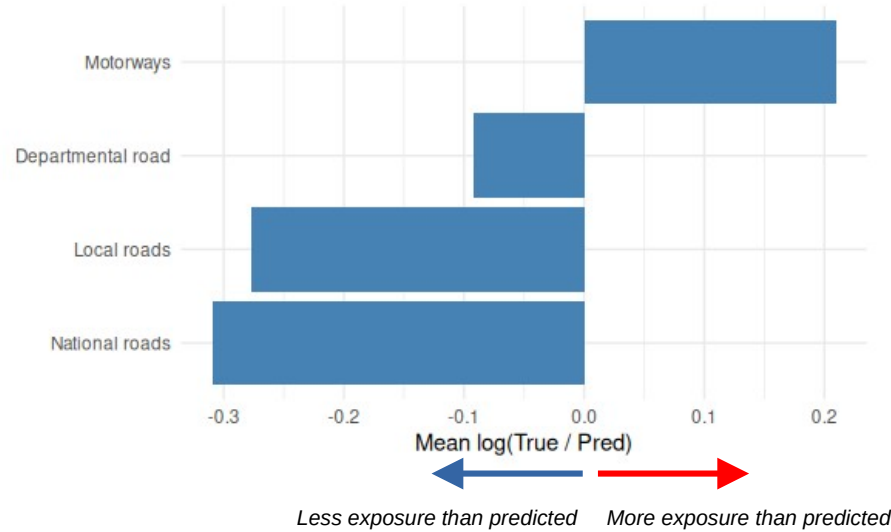
Correlation (Pearson) = 0.19
Correlation (Spearman) = 0.12
RMSE = 2.63
MAE = 0.66



H1: Are road visits by participants explained by road traffic?

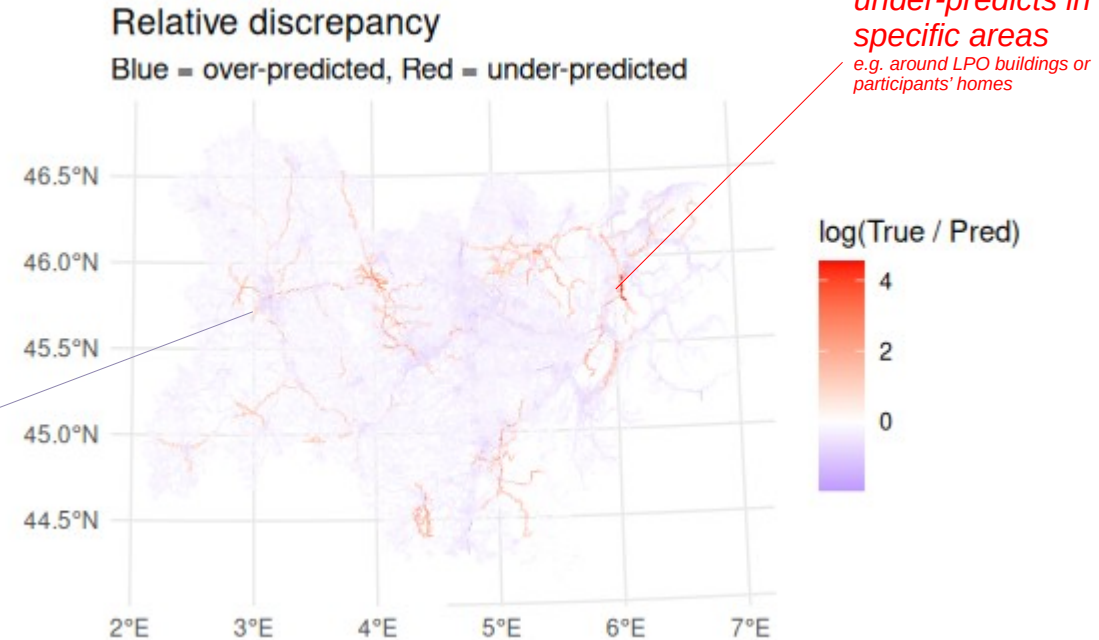


H1: Are road visits by participants explained by road traffic?

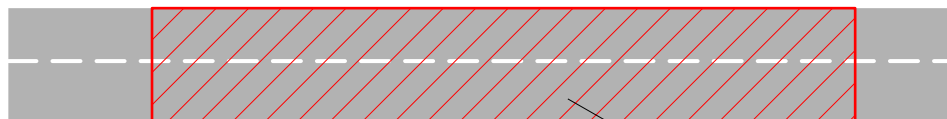


→ *AADT proxy assumes participants are using the road network like the average driver*

*over-predicts in cities
(where participants
don't live/drive)*

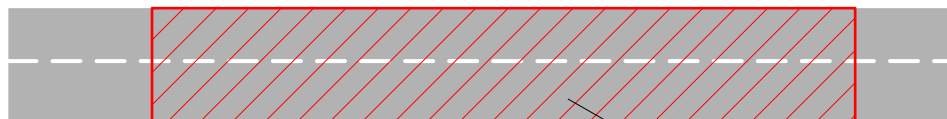


H2: Can roadkill counts capture sampling effort?



1 visit = 1 or more reports / journey

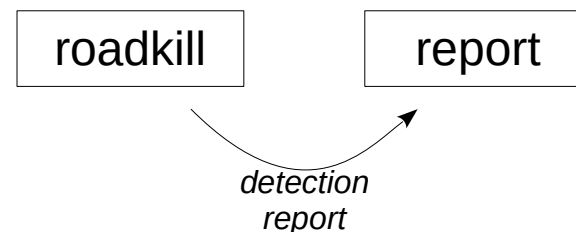
H2: Can roadkill counts capture sampling effort?



1 visit = 1 or more reports / journey

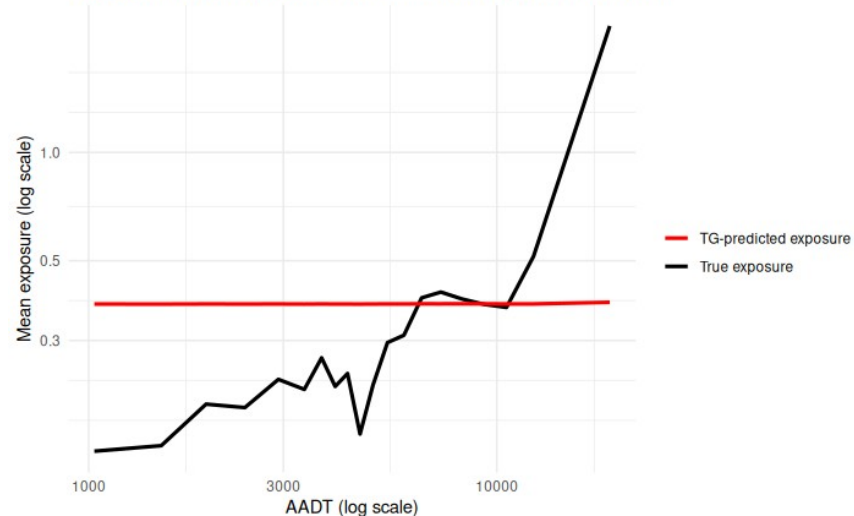
419 reports during the experiment
→ 0.05 reports/km (low rate)

Likely more reports than normal because experiment increased focus & willingness to report



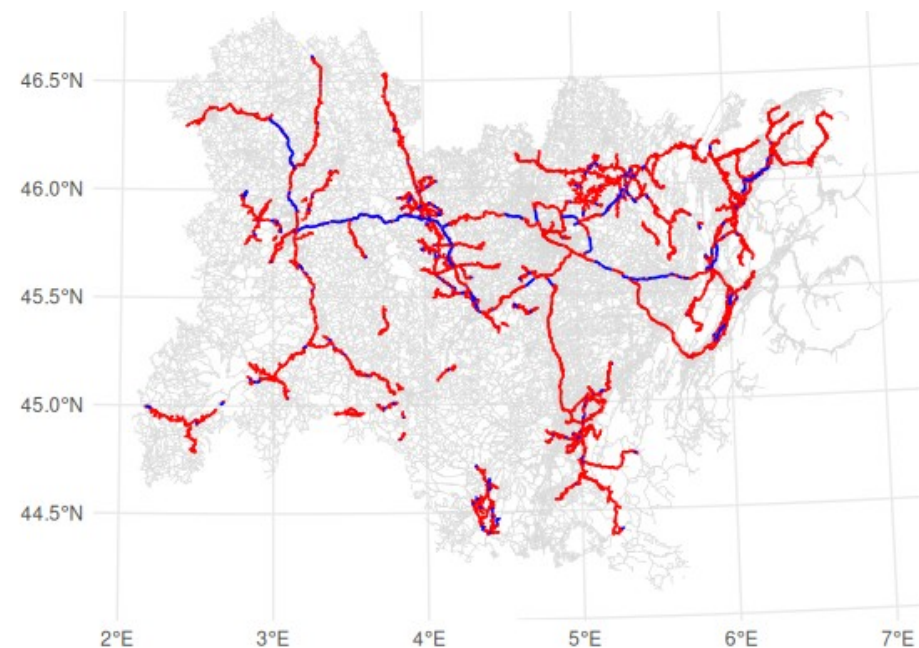
H2: Can roadkill counts capture sampling effort?

Approximate road exposure with Target-Background methods



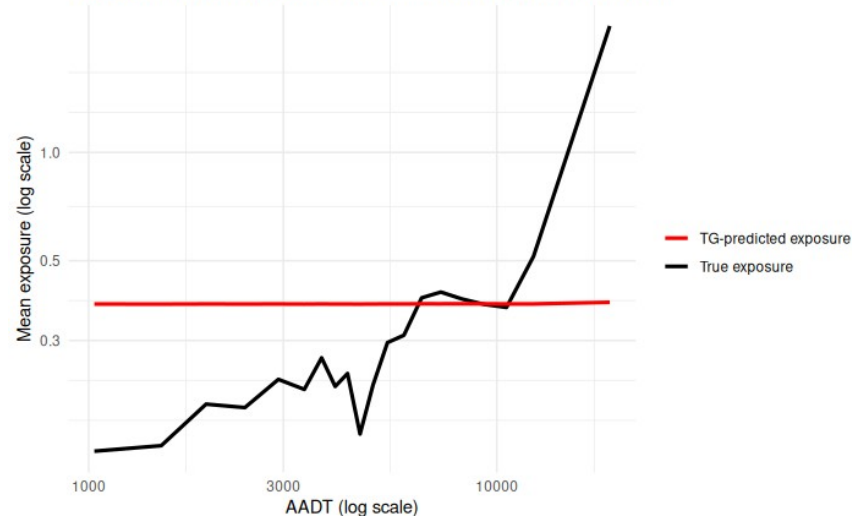
Correlation (Pearson) = 0.16

Correlation (Spearman) = 0.15



H2: Can roadkill counts capture sampling effort?

Approximate road exposure with Target-Background methods



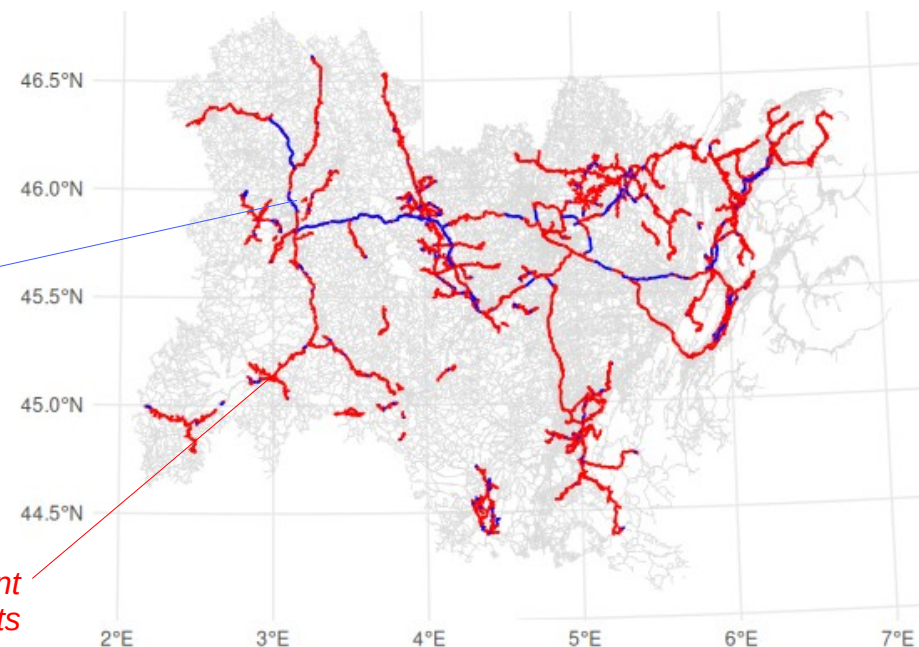
→ *TG on roadkill is not viable*

Contrary to living species, roadkill distribution is never neutral enough (even with all species combined)
→ *can't erase the ecological process to inform reporting rates*

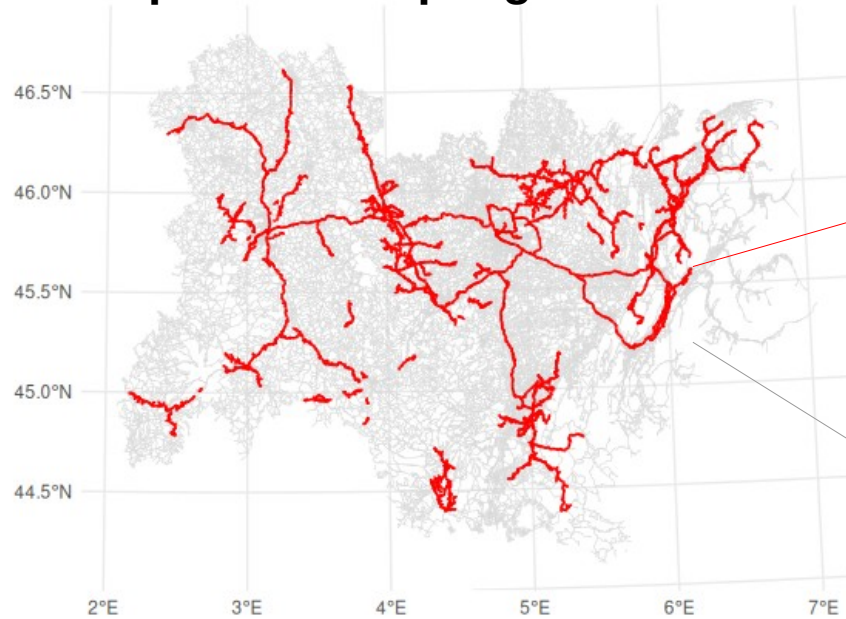
Roadkill is scarce, + citizen science = low detection, low reporting rates
→ *lots of roads are visited often but have nothing to show for it*

Segment with a report

Visited segment with no reports



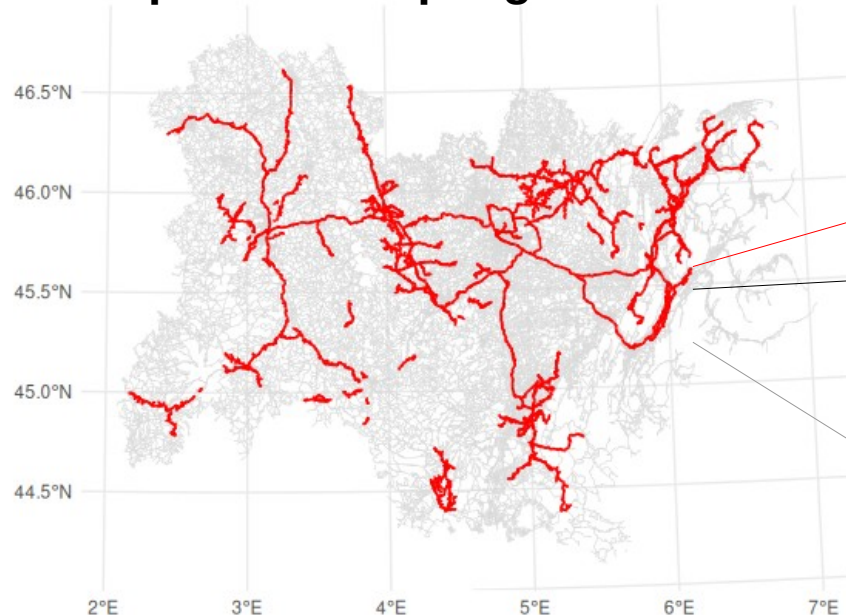
Can we predict sampling effort at all?



Roadkill reports (TG): proof of visit, but highly incomplete

AADT = proxy for road usage everywhere, but ignores actual participants habits

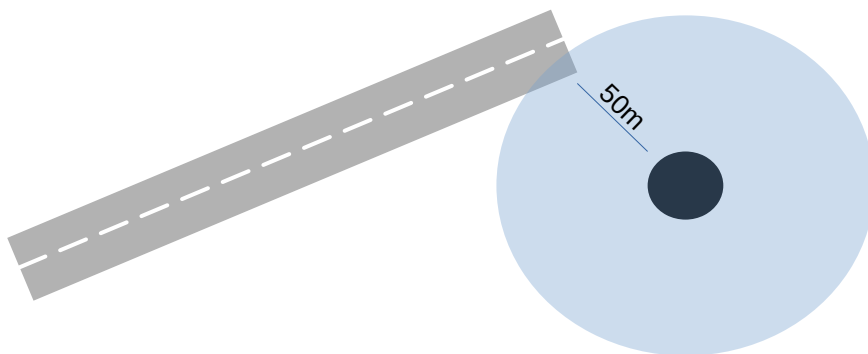
Can we predict sampling effort at all?



Roadkill reports (TG): proof of visit, but highly incomplete

Living reports (Living TG) can help identify the journey's start and endpoints
→ assuming people travel mostly by car

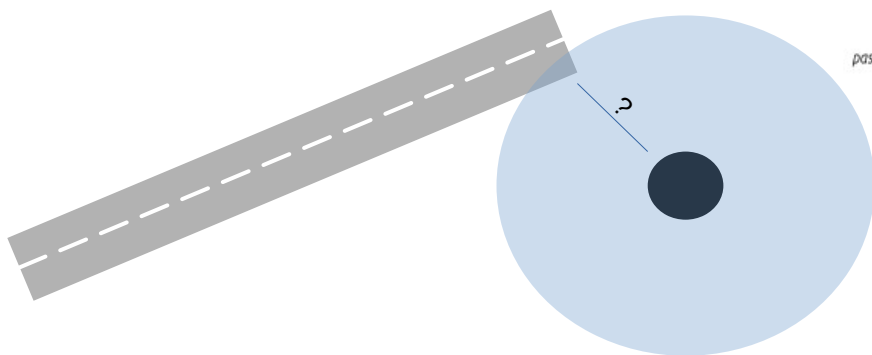
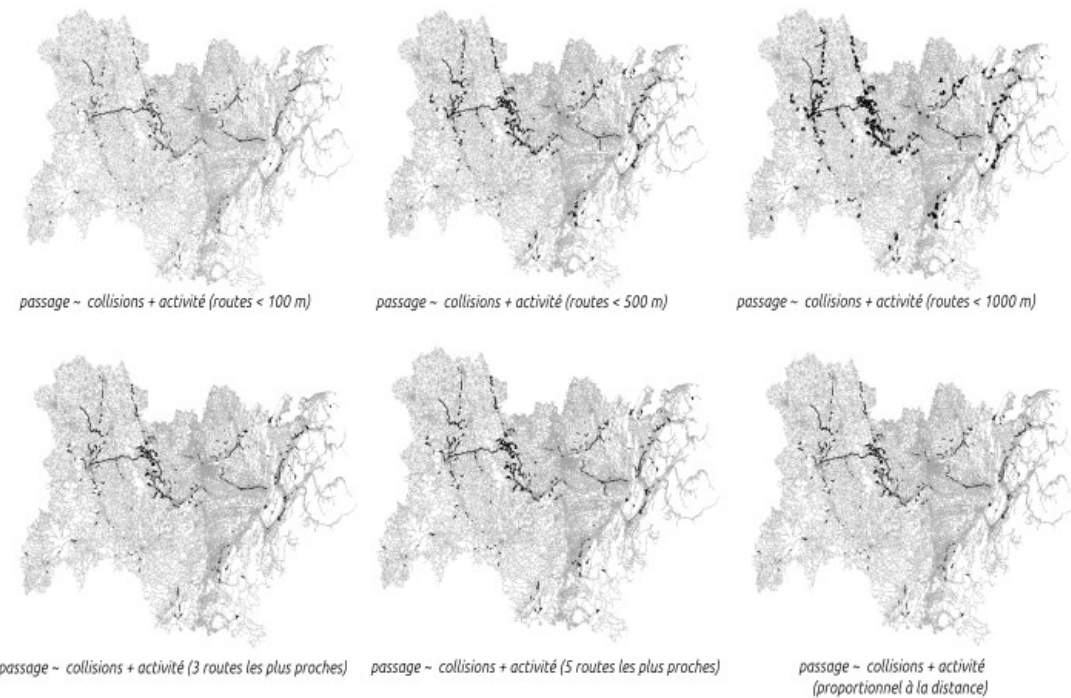
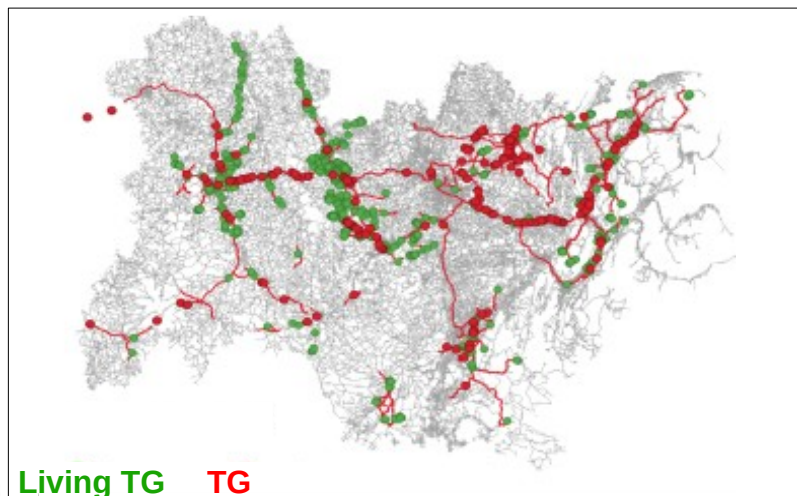
AADT = proxy for road usage everywhere, but ignores actual participants habits



Living reports:

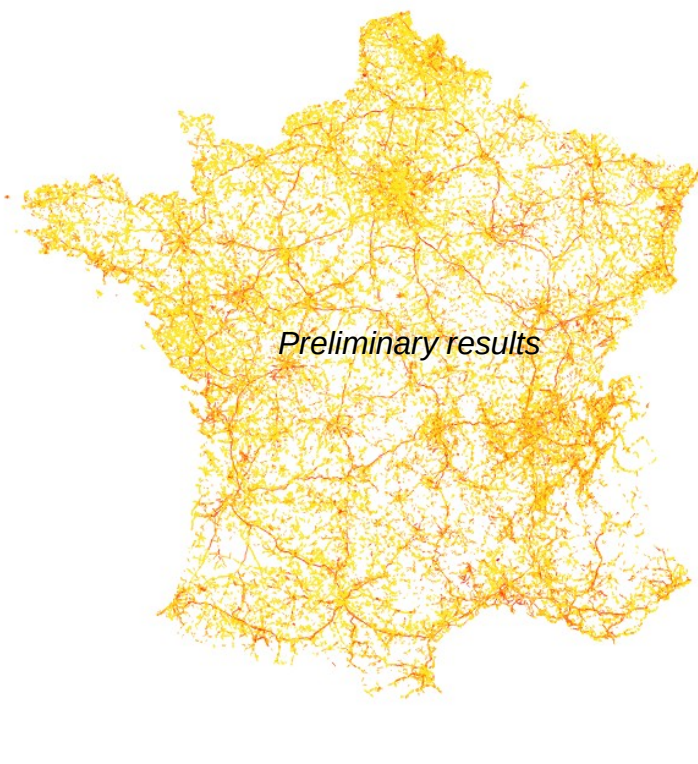
→ The activity of a *roadkill data participant* in reporting living animals to Faune France

Can we predict sampling effort at all?



Can we predict sampling effort at all?

We can combine all 3 complimentary approaches to predict road exposure in Faune-AuRA



Random forest model

- living TG (KDE-based propagation)
- roadkill TG
- AADT

→ Pure regression (number of visits) is too difficult
($R^2 = 0.07$)

→ Broad categories of visit frequency per road could be achievable (F1 score = 0.53)



Conclusions

1. Traffic volume is not a reliable proxy for participant effort

- Intuitive choice, but did not reflect where Faune-AuRA contributors actually travel: professionals work in remote areas, relatively few contributors live or commute in cities
- Traffic is itself a driver of roadkill

Conclusions

1. Traffic volume is not a reliable proxy for participant effort

- Intuitive choice, but did not reflect where Faune-AuRA contributors actually travel: professionals work in remote areas, relatively few contributors live or commute in cities
- Traffic is itself a driver of roadkill

2. Roadkill data alone is too sparse and clustered for a Target-Group approach

- Roadkill is always clustered
- Reports are sparse (low detection + low reporting rates)
 - *Living* reports: complementary information, partially filling the gaps

—► **Predicting road visits in the Faune-France dataset remains challenging**

Transmorta included varied participant profiles (incl. the top users in the region) *but* still only 30 active & occasional users out of 2,900 registered

→ *For GDPR and other legal reasons, tracking all vehicle movements is not feasible for a random subset of FF participants*

Conclusions

1. Traffic volume is not a reliable proxy for participant effort

- Intuitive choice, but did not reflect where Faune-AuRA contributors actually travel: professionals work in remote areas, relatively few contributors live or commute in cities
- Traffic is itself a driver of roadkill

2. Roadkill data alone is too sparse and clustered for a Target-Group approach

- Roadkill is always clustered
- Reports are sparse (low detection + low reporting rates)
- *Living* reports: complementary information, partially filling the gaps

—► **Predicting road visits in the Faune-France dataset remains challenging**

Transmorta included varied and representative participant profiles (incl. the top users in the region) *but* still only 30 active users out of 2,900 registered

→ For GDPR and other legal reasons, tracking all vehicle movements is not feasible for a randomly selected sample of FF participants

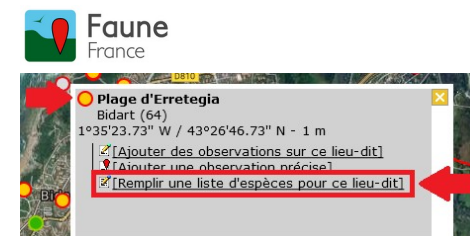
Perspectives

Report species in driving transects (list-based reporting)?

- Available in FF but other projects report low adoption for roadkill data
- Latent spatial biases persist: high driving speed reduces detection; difficult parking reduces reporting...

Joint modelling of structured and opportunistic data?

- Standardized roadkill surveys (authorities, insurance data, patrols, academic surveys...) together with opportunistic observations



Merci !

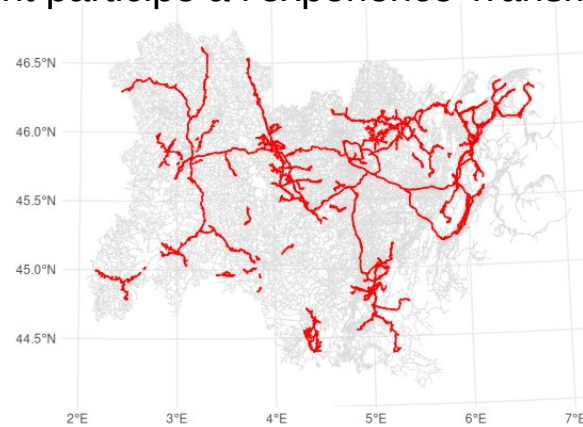


MINISTÈRE
DE LA TRANSITION
ÉCOLOGIQUE,
DE L'ÉNERGIE, DU CLIMAT,
ET DE LA PRÉVENTION
DES RISQUES

Liberté
Égalité
Fraternité



Un grand merci aux
salariés et adhérents de la LPO AuRA
qui ont participé à l'expérience *Transmorta*



PRÉFET
DE LA RÉGION
AUVERGNE-
RHÔNE-ALPES

Liberté
Égalité
Fraternité

FONDATION
FRANÇOIS
SOMMER

