RTM: a technical exploration

To better understand how RTM works, we put to ourselves the goal of building a toy model. RTMs can be complex, and many of them, such as the one used by the Metropolitan Police of London, are proprietary. This means that information about its functionality is limited to self-disclosure. However, RTM generally functions following the same overall functioning that we tried to simulate. For the sake of convenience, we decided to try to build a simple RTM for the city of Paris.

A first (failed) attempt

RTM works around a target variable, which is what the model should predict. To do so, a set of "risk factors" are compiled and analysed, using generally either a Poisson regression (as in Steif, 2021) or, if dispersion of data is too wide, a Negative Binomial Regression.

For our target variable, the first, most natural element, seemed to be data on criminality that the French Ministry of Interior publishes and updates regularly in open data (Ministère de l'Intérieur, 2024). Data is available at municipal level for the entire French territory, for a number of offences (that exclude murders and attempted murders). For Paris, the finest granularity level is the arrondissement level.

For a first exploratory analysis, we decided to use, among the multiple crimes, drug trafficking. For our risk factors, we used data, available in the same open format provided by the French public administration, on the urban issues signaled through the app "Dans Ma Rue", and the location of Tabac stores. The idea, somehow naive, was that a correlation might arise between drug trafficking and the commerce of another addictive substance (tobacco), as well as with urban degradation.

To combine the risk factors and the target variable, we first used the Poisson regression, offsetting by population at the arrondissement level. This regression showed a significant correlation between the presence of tobacco shops, urban signals and drug trafficking. Both variables had three "*" next to them. However, at a closer look, it was evident that the level of dispersion was such to make the model practically useless.

> This brought us to a first reflection: if a model is proprietary, a decision (conscious or not) could be made to only show a part of the model's results to members of the police. Moreover, which statistical competences does the police have at its disposal to evaluate the statistical integrity of a Poisson regression model?
>
> This was also a practical representation of what is called technical bias. We had been very strict on the idea of not using sensitive socio-economic data or proxies, but our model was skewed (against arrondissement with more urban issues or tobacco shops) for technical reasons. Conversations on data in the public sector sometimes fail to address the simple need for rigorous statistics.

At this point, we switched to a Negative Binomial Generalized Linear Model (GLM). This model is a better fit for cases of high dispersion, as the one in our hands. As a matter of fact, this model showed no correlation between the risk factors and the target variable.

We soon realised that, while an interesting experiment, the model we had built was far from realistic. There were two main reasons, first of all RTM is supposed to be a "close" look on a city, and the examples that we had found online used grids of a couple of hundreds of meters - our model used much much larger containers, just 20 arrondissements. This was due to the impossibility of accessing more detailed data on crime in the city. Moreover, we were not able to come up with more relevant risk factors for the target variable we had decided to investigate.

A second attempt

We decided to start from scratch, and make the reports on Dans Ma Rue our target variable. The R script, data used and visualisation of this section are available at the following link: https://github.com/giocerb/RTM-simulation. It was a bit less exciting to anticipate trivial complaints in the city, rather than drug trafficking, but the data available was abundant and granular. Nevertheless, the investigation maintained its relevance for public policy and the adaptation was functional to make us better understand how the Metropolitan Police's system might actually work.

We started by importing data from Insee of France as a 200mx200m grid. It was a complex task, and perhaps a similar result could have been achieved in a more economical way, but it was definitely an opportunity to practice how different coordination systems interplay and can be manipulated in the R ecosystem.

For our risk factors, we decided to obtain data on the distribution of clubs, fast food locations, metro stops and police stations. The rationale was that more public dissatisfaction could cluster around night life spots and places that are often busy, such as metro stops and fast food restaurants. We also imagined that around police stations, civic action could be more quickly prompted by the city government itself - without the intervention of citizens.

We layered these factors on the grid, calculating the density of each of them per unit of the grid.

We also created a function to calculate the distance between each grid to the closest of the elements.

Having done this, we fitted a series of negative binomial generalized linear models (GLMs), comparing the explanatory power of density-based variables, simple counts, and nearest-neighbour distances. Across all specifications, the model based on distance (with k = 1 nearest neighbours) emerged as the strongest performer. It produced the lowest AIC, the largest reduction in deviance from the null model, and the most consistent and statistically significant coefficients.

Statistical results

Predictor	**Estimate (β)**	**IRR (eβ)**	**Std. Error**	**z-value**	**P-value**
Intercept	10.576	39,188.0	0.035	301.15	< 0.001 ***

Dist. to Clubs	-0.348	**0.706**	0.029	-12.07	< 0.001 ***
Dist. to Metro Stops	-1.163	**0.313**	0.119	-9.75	< 0.001 ***
Dist. to Fast Food	-1.994	**0.136**	0.144	-13.86	< 0.001 ***
Dist. to Police	-0.295	**0.745**	0.045	-6.55	< 0.001 ***

Model Statistics:

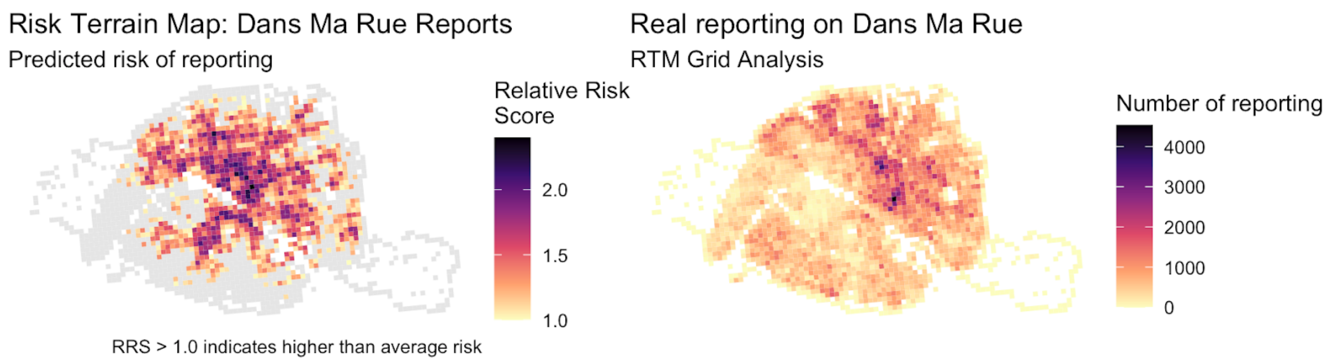
- **Observations**: 2,198
- **Theta**: 1.369 (SE: 0.040)
- **AIC**: 31,087
- **Log Likelihood:** -15,537.28

The distance model suggested a clear spatial gradient: grid cells located closer to metro stations, nightlife venues, fast-food outlets, and police stations exhibited higher levels of reported issues on *Dans Ma Rue*.

Limitation, conclusions and AI use

This exploration was a useful exercise to better understand the inner workings of RTMs, as the one the Metropolitan Police.
Data and algorithms in the public sector, as we argue across this report, have significant issues of opacity. Opacity comes from the complexity of its statistical modelling, either intrinsic or proper of the observer of the model, as well from intellectual property protections. Our work contributes to both of these dimensions, as we provide a simple model built from scratch covering roughly the entirety of the model life (from data handling to visualisation) using open data and free software (R and its libraries).

However, there are obvious and evident limitations to our work. The data used could have been more specific, and its handling suffers from a lack of statistical knowledge. We explored moran I test, validation and residual methods, but we felt like excluding them from our discussion due to our limited understanding of them.
Our resources were a book by Ken Steif (2021), and our work was made in co-creation with various LLMs, mostly Gemini's 3 pro and Claude Sonnet 4.5. They were used to debug and find specific solutions, interpret results, and clean the final script. A high standard of understanding of the LLMs output was pursued throughout the work.



The toy model can be found at the following repository:
https://github.com/giocerb/RTM-simulation