

---

**TROIS PETITS PAPIERS EN QUÊTE D'AUTEUR**

Luigi PIRANDELLO (\*)

(\*) Insee, Direction de la méthodologie et de la coordination statistique et internationale

[luigi.pirandello@insee.fr](mailto:luigi.pirandello@insee.fr)

**Mots-clés** : Covid, évolution temporelle, distance entre lois de probabilités, décalage temporel, statistique descriptive

**Domaine concerné** : Statistique descriptive, analyse des séries temporelles, impact de la Covid

---

### Résumé

**Papier n°2 : comparaison des profils temporels d'évolution d'une épidémie entre deux zones géographiques**

La pandémie de Covid-19 a frappé le monde et la France en particulier depuis le début de l'année 2020, avec des évolutions temporelles très marquées (identification de plusieurs « vagues » de contaminations, entrecoupées de périodes de régression) et dans une certaine mesure assez différentes d'une zone géographique à une autre, avec des profils pouvant mettre en évidence des décalages temporels.

Quels que soient les indicateurs que l'on retienne (et l'on peut se concentrer sur les plus objectivables, les moins entachés d'erreurs : nombre de morts journaliers, nombre d'entrées en soins intensifs, nombre de malades hospitalisés un jour donné), on observe ces deux phénomènes : profils temporels plus ou moins cycliques, disparités régionales.

**Ce papier se propose de fournir une méthode de comparaison des profils d'évolution entre deux zones géographiques.** L'optique est ici purement descriptive, elle ne vise pas à proposer des projections dans le futur ni des estimations de paramètres décrivant l'évolution de l'épidémie. Principalement, la méthode suggérée doit mettre en évidence les décalages temporels dans la propagation du fléau entre deux zones ou catégories de population.

Dans un premier temps, on examine comment comparer deux distributions de probabilités entières et *tronquées à droite*, pour représenter les observations dont on dispose à un instant T pour différents domaines, alors que la pandémie n'est pas achevée et que les profils d'évolution observés sont tronqués.

Ensuite, on examine si une nouvelle distribution peut être considérée comme issue d'une distribution de référence mais décalée dans le temps. On utilise pour cela des indicateurs de similarité tels que la *distance du KHI2* ou *l'information de KULLBACK symétrisées*, notées  $\Delta(i)$ .

Dans les deux cas, on ne connaît pas le paramètre de décalage  $i$ , qu'il faut donc estimer. On peut obtenir une estimation en cherchant la valeur  $\hat{i}$  réalisant  $\text{Min } \Delta(i)$  pour  $i$  variant de 0 à T. L'indicateur de distance entre les deux distributions (celle de référence et celle potentiellement décalée) sera donc  $\Delta(\hat{i})$ .

### Mise en œuvre pratique

Dans la pratique, on observe les valeurs journalières d'une variable d'intérêt pour un département ou une zone géographique donnée, par exemple le nombre  $n_k$  de décès ou le nombre de nouvelles contaminations par la Covid-19, comptés à partir de l'instant 0 par convention.

On note :  $N(T) = \sum_{j=0}^T n_j$ , où T est la date finale des observations disponibles.

Pour un autre département ou une autre zone, on dispose d'observations similaires notées respectivement  $n'$  et  $N'$ .

On cherche à savoir si l'évolution de la maladie dans les deux départements ou zones est similaire, avec éventuellement un décalage temporel  $i$ , dont on cherche une estimation.

*Nota :*

Si certains  $n_k$  ou  $n'_k$  sont nuls, on agrège les observations, par exemple avec celles des instants suivants ou on lisse par moyenne mobile courte, pour obtenir des valeurs non nulles. Bien entendu, les distributions empiriques observées peuvent être nulles définitivement à partir d'une certaine date (cas de l'extinction du phénomène considéré).

### *Statistique descriptive*

On peut classer les couples de départements par valeurs décroissantes de l'indicateur de distance ci-dessus en indiquant la valeur du décalage estimé  $\hat{i}$ , ce qui mettra en évidence les similitudes ou les dissemblances entre les départements ou zones, en termes d'évolution temporelle de la variable d'intérêt mais indépendamment du niveau de celle-ci.

Il faut prendre garde, cependant, à ne pas faire de comparaisons sur des périodes trop courtes, avec un nombre trop faible d'observations, ce qui aurait pour effet d'accepter l'hypothèse de similarité. Pour la même raison, la plage de balayage du paramètre  $i$  de décalage temporel doit être bornée empiriquement pour éviter des valeurs trop fortes conduisant à un recouvrement des séries trop faible.

Dans un second temps, on pourrait chercher à constituer des groupes de zones homogènes au vu de leurs profils temporels d'évolution.

***Ces petits papiers cherchent un (co-)auteur pour mettre en application sur des données réelles les méthodes exposées, tester leur pertinence et apporter tout complément utile....***

### Bibliographie