Journées de méthodologie statistique de l'Insee 2022

TROIS PETITS PAPIERS EN QUÊTE D'AUTEUR

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

Marc CHRISTINE (*)

(Luigi PIRANDELLO[1])

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

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

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

[1] Ce papier a été présenté aux JMS2022 sous le pseudonyme de Luigi PIRANDELLO.

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 la *distance du KHI2 ou l'information de KULLBACK symétrisées, notées Δ(i).*

Short abstracts in English and French

This paper focuses on methods for comparing the temporal evolution of a disease (such as Covid-19) between two geographical areas. The method should mainly highlight similarities in the evolution with possible time lags. The paper proposes simple methods to evaluate the gap between two temporal distributions and to estimate the lag, based on the KHI2 distance or the KULLBACK information. As an application, one can consider using descriptive statistics to compare two zones according to their greater or lesser similarity. We can also try to create homogeneous groups of zones using ascending hierarchical classification methods.

Dans ce papier, on s'intéresse aux méthodes de comparaison des profils d'évolution temporelle d'une maladie (telle que la Covid-19) entre deux zones géographiques. La méthode doit principalement mettre en évidence des similitudes des évolutions avec d'éventuels décalages temporels. Le papier propose des méthodes simples pour évaluer l'écart entre deux distributions temporelles et estimer le décalage, fondées sur la distance du KHI2 ou l'information de KULLBACK.

À titre d'application, on peut envisager de faire de la statistique descriptive pour comparer les zones deux à deux selon leur plus ou moins grande similarité. On peut aussi chercher à constituer des groupes homogènes de zones à l'aide de méthodes de classification hiérarchique ascendante.

1. Cadre théorique

a. Comparaison de deux distributions entières avec décalage exact

Soit X une variable aléatoire entière, ≥ 0, de loi définie par p_k= P{X = *k*} ≠ 0, de fonction de répartition :

$$F(q) = P\{X \leq q\} = \sum_{j=0}^{q} p_j \ .$$

On définit Y = X + *i* (*i* ≥ 0). La loi de Y est donnée par q_k = P{Y = *k*} = P{X = *k* - *i*}, soit :

$$q_k = \begin{cases} 0 \ si \ k < i \\ p_{k-i} \ sinon \end{cases}.$$

b. Cas d'une troncature à droite

Dans la pratique, on ne peut pas observer toutes les valeurs de X ni de Y mais seulement celles qui sont inférieures ou égales à T. Ceci conduit à travailler sur les *distributions conditionnelles* de X et de Y sachant X (resp. Y) ≤ T.

On note : $p'_k(T)$ = P{X = *k* / X ≤ T} = P{X = *k* et X ≤ T} / P{X ≤ T}, soit :

$$p'_k(T) = \begin{cases} 0 \ si \ k > T \\ \dfrac{p_k}{F(T)} \ sinon \end{cases}.$$

De même pour Y : $q'_k(T)$ = P{Y = *k* / Y ≤ T} = P{X = *k* - *i* et X ≤ T - *i*} / P{X ≤ T - *i*} (défini si T – *i* ≥ 0), soit :

$$q'_k(T) = \begin{cases} 0, si \ k > T \\ \dfrac{p_{k-i}}{F(T-i)} \ sinon \ (= 0 \ si \ k < i) \end{cases}.$$

Dans la suite, on imposera les conditions : *i* ≤ *k* ≤ T pour éviter les cas d'annulation des termes intervenant dans les formules ci-dessus.

c. Comparaison de deux distributions.

Soit Z une autre variable aléatoire entière, de loi définie par $P\{Z = k\} = r(k) \geq 0$ et de fonction de répartition H : $H(t) = P\{Z \leq t\}$. On observe seulement les valeurs de X et de Z inférieures ou égales à T, et on cherche à voir si Z peut être considérée comme décalée de X de la quantité *i*.

Pour cela, on va comparer les *distributions conditionnelles* de X et de Z sachant X (resp. Z) ≤ T. On note :

$$r'_k(T) = P\{Z = k \ / \ Z \leq T\} = r(k) \ / \ H(T).$$

Si Z est exactement de la forme X + *i*, on aura : $r'_k(T) = \frac{p_{k-i}}{F(T-i)}$ pour *i* ≤ *k* ≤ T.

► Pour analyser dans quelle mesure Z peut être approchée par X ***+ i***, on va utiliser un indicateur d'écart entre la loi effective de Z sachant Z ≤ T et la loi théorique correspondant au cas d'un décalage exact.

- *distance du KHI2 symétrisée*

Cette distance est donnée par :

$$\Delta(i) = \sum_{k=i}^{T} \left[\frac{r_k}{H(T)} - \frac{p_{k-i}}{F(T-i)}\right]^2 \left[\frac{F(T-i)}{p_{k-i}} + \frac{H(T)}{r_k}\right].$$

Dans la pratique on peut se contenter de supposer que $r_k > 0$ seulement pour $k \geq i$, pour que l'expression ci-dessus ait un sens.

- information de KULLBACK symétrisée :

Elle est donnée par :

$$\boldsymbol{K(i)} = \sum_{k=i}^{T} Ln\left[\frac{p_{k-i}}{F(T-i)} \Big/ \frac{r_k}{H(T)}\right]\left[\frac{p_{k-i}}{F(T-i)} - \frac{r_k}{H(T)}\right]$$

d. Identification et estimation d'un décalage éventuel entre les distributions

L'un des paramètres important est le décalage *i*. On peut l'estimer en cherchant la valeur *î* réalisant Min Δ(*i*) pour *i* variant de 0 à T. L'indicateur de distance entre les deux distributions (celle de référence correspondant à X et celle potentiellement décalée correspondant à Z) sera donc Δ(*î*).

2. Mise en œuvre pratique

a. Contexte d'application

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 d'un instant initial 0.

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'_k et $N'(T)$.

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

b. Mise en œuvre (cas de la distance du KHI2)

On va donc, conformément au modèle théorique ci-dessus, chercher :

$$Min_{0\leq i\leq T}\ \Delta(i) = \sum_{k=i}^{T} \left[\frac{n'_k}{N'(T)} - \frac{n_{k-i}}{N(T-i)}\right]^2 \left[\frac{N(T-i)}{n_{k-i}} + \frac{N'(T)}{n'_k}\right].$$

Dans la pratique, si certains n_k ou n'_k sont nuls, on agrège les observations, par exemple avec celles des instants suivants, pour obtenir des valeurs non nulles. On peut aussi choisir de ne traiter les données qu'en *cumul* (par exemple hebdomadaire), voire opérer un *lissage* préalable.

Bien entendu, les distributions empiriques observées peuvent être nulles définitivement à partir d'une certaine date (cas de l'extinction du phénomène considéré).

c. 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 *î*, 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 de conduire à accepter plus fréquemment 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.

d. Constitution de groupes homogènes

Une seconde application viserait à constituer des groupes de zones homogènes de départements ou zones géographiques au vu de leurs profils temporels d'évolution, selon des méthodes de classification ascendante hiérarchique avec contraintes de contiguïté telles que développées dans [1] et [2].

3. Conclusion

Comme chez l'auteur de la pièce[2], ces petits papiers cherchent un (co-)auteur pour continuer l'histoire, mettre en application les méthodes exposées sur des données réelles, tester leur pertinence et apporter tout complément utile….

Bibliographie

[1] Christine M., Isnard M., "Agrégation optimale sous contraintes de contiguïté : aspects théoriques et mise en oeuvre avec applications à des cas pratiques", Actes des 11èmes Journées de méthodologie statistique de l'Insee (2012) sur http://jms-insee.fr
[2] Christine M., Isnard M., "Aurait-on pu construire des régions selon des critères statistiques ? ", Actes des 12èmes Journées de méthodologie statistique de l'Insee (2015) sur http://jms-insee.fr

[2] Luigi PIRANDELLO : « *Sei personnagi in cerca d'autore* » (1921)