Méthodologie – Ressemblances

Description générale

Les graphiques de ressemblance permettent de visualiser pour un indicateur et plus[1] les régions les plus proches statistiquement d'une région de référence.

La proximité statistique est sensible à la méthode de mesure utilisée. Regioviz propose deux méthodes classiques et facilement interprétables. La **distance euclidienne** (ou distance à vol d'oiseau) prend en compte des données préalablement standardisées. Si la valeur de l'indice équivaut à 0, la similarité est totale entre ces deux unités territoriales. Plus la valeur de la distance est élevée, moins la similarité est importante. **L'écart de rang moyen** évalue la distance, exprimée en rang, qui sépare deux unités territoriales. La valeur de cet indice est théoriquement comprise entre 1 (écart de rang minimal) et le nombre d'unités territoriales que constitue l'espace d'étude – 1 (écart de rang maximal).

L'interface Regioviz propose deux niveaux pour la visualisation de ces ressemblances : la **ressemblance globale** et la **ressemblance détaillée** indicateur par indicateur.

L'option de distance globale propose une visualisation synthétique de l'éloignement statistique existant entre « ma région » et les autres régions de l'espace d'étude sur les *n* indicateurs sélectionnés. Ce module est composé d'un graphique en essaim (beeswarm) qui permet de visualiser graphiquement le degré de ressemblance statistique existant entre ma région et les autres régions de l'espace d'étude. La carte associée à la représentation graphique rend compte de l'organisation spatiale de ces proximités statistiques : les 25 % des indices de similarité les plus faibles (régions les plus ressemblantes) apparaissent dans des tonalités rouges, les 25 % les plus importantes (régions les moins ressemblantes) sont représentées par des tonalités bleues.

Pour comprendre quel est le poids de chaque indicateur dans la mesure de ressemblance globale, Regioviz propose systématiquement une représentation graphique permettant d'évaluer visuellement le degré de similarité indicateur par indicateur (ressemblances par indicateur). Par défaut, l'application décompose cette ressemblance pour l'unité territoriale qui ressemble le plus à « mon territoire » de référence d'après la mesure globale de ressemblance. Libre ensuite à l'utilisateur de choisir plus ou moins d'unités territoriales de comparaison (les *n* unités les plus ressemblantes) en fonction de ses objectifs d'analyse.

Mesure de distance entre régions

Le calcul de distance entre plusieurs individus statistiques pour plusieurs variables quantitatives est un sujet courant en statistiques et constitue notamment une étape préalable aux classifications et mesures de ressemblances. La mesure de la distance entre deux points peut être définie de plusieurs manières en fonction des objectifs de l'analyse.

Pour Regioviz, la première méthode est la **distance euclidienne**. Couramment utilisée et intuitive, elle correspond à la distance la plus courte dans un espace homogène et isotrope. Néanmoins, la prise en compte de variables aux unités de mesure et ordres de grandeur hétérogènes tend à affecter plus de poids aux variables caractérisées par une forte dispersion, neutralisant presque complètement l'effet des autres variables. Pour ne pas privilégier les variables caractérisées par une forte dispersion, celles-ci sont préalablement normalisées : elles sont centrées par la moyenne et réduites par l'écart type (standardisation).

La distance euclidienne normalisée entre deux régions utilisée dans Regioviz s'écrit alors :

$$\text{Dist}\left(X_i, X_{i'}, j, p\right) = \sqrt{\sum_{j=1}^{p}\left(\frac{X_i^j - \bar{X}^j}{\sigma^j} - \frac{X_{i'}^j - \bar{X}^j}{\sigma^j}\right)^2}$$

Avec :
- X_i : Ma région
- $X_{i'}$: Une région i' à comparer à ma région
- p : Le nombre p d'indicateurs sélectionnés
- $\bar{x}^j$: Moyenne pour l'indicateur j
- σ^j : Écart-type pour l'indicateur j

[1] Pour des raisons de lisibilité sur le graphique, le maximum est fixé à 7 indicateurs.

La seconde méthode est la **l'écart de rang moyen**. En comparaison à la précédente mesure de distance, cette méthode ne prend pas en compte la distance statistique entre deux objets statistiques. Elle présente néanmoins l'avantage d'être facilement interprétable (« 3,4 rangs séparent en moyenne mon territoire avec cet autre territoire pour les *p* indicateurs sélectionnés ») et d'être nettement moins sensible aux valeurs extrêmes lors de la visualisation du graphique dans Regioviz que la distance euclidienne. Enfin, la confrontation de méthode apporte toujours de la profondeur d'interprétation puisqu'elle rappelle que la mesure d'une distance est toujours dépendante des paramètres considérés pour sa mesure.

Cet écart de rang moyen correspond à la moyenne des écarts en valeur absolue entre une région et une autre. Cette distance de rang entre deux régions utilisée dans Regioviz s'écrit alors :

$$\text{Dist}\left(X_i\ X_{i'\ j}\ p\right) = \frac{\sum_{j=1}^{p} \left|Rank_i^j - Rank_{i'}^j\right|}{p}$$

Avec :

X_i : Ma région
$X_{i'}$: Une région i' à comparer à ma région
p : Le nombre p d'indicateurs sélectionnés
Rank : La position (rang) dans une distribution donnée

Description des options de sélection

1 L'utilisateur est invité à choisir sa méthode de mesure des ressemblances. Les valeurs de distance sont exprimées en abscisses. Avec la méthode de calcul de distance euclidienne, plus le nombre d'indicateurs sélectionnés est important, plus les valeurs de distance augmentent. Ma région apparait en jaune à gauche du graphique (distance de 0). La région la plus semblable est celle qui est située immédiatement à droite de celle-ci. La région la moins ressemblante est celle figurant à l'extrême droite du graphique.

2 La mesure de ressemblance s'applique aux régions sans valeurs manquantes. Les couleurs symbolisent les niveaux de ressemblance globale : région la plus ressemblante (rouge foncé), puis les 25 % les plus ressemblantes (régions très ressemblantes, rouge), les 25 % à 50 % les plus ressemblantes (régions ressemblantes, orange), les 50 % à 75 % les plus ressemblantes (régions dissemblantes, bleu clair) et enfin les 75 % à 100 % les plus ressemblantes (régions très dissemblantes, bleu foncé).

3 Il est possible d'identifier la localisation sur la carte de ces points grâce à un rectangle de sélection. De la même façon il est possible de sélectionner sur la carte une ou plusieurs régions et visualiser leurs positions sur le graphique.

4 En cliquant sur l'option « ressemblances par indicateur », une fenêtre détaillée par indicateur composant l'indice de ressemblance global apparait. Par défaut ma région apparait au centre des valeurs axe par axe et la région la plus proche est représentée pour comparer ses valeurs avec celle de ma région.

5 Par défaut, les régions ressemblantes qui sont caractérisées par des valeurs supérieures à ma région sont représentées en rouge ; et celles par des valeurs inférieures en vert. En cliquant sur le « + » ces couleurs sont inversées.

6 Il est possible d'inverser l'ordre d'apparition des axes en cliquant sur les flèches des axes. Cette option peut s'avérer utile pour rapprocher les axes qui présentent des similitudes statistiques ou thématiques.

7 Par défaut, la taille des cercles est constante. En activant l'option « cercles proportionnels à la population » la taille des cercles est dorénavant proportionnelle à la population des régions. Cette option peut s'avérer utile pour nuancer le positionnement des régions en fonction de leurs poids respectifs.

8 Par défaut, n'est représentée sur le graphique que la région la plus proche sur l'indice de ressemblance global. Mais L'utilisateur peut choisir de mettre en évidence le positionnement statistique de la $n^{ème}$ région la plus proche. Dans ce cas de figure, les régions globalement plus ressemblantes que cette $n^{ème}$ région sont représentées en vert et rouge sur le graphique, mais avec une légère transparence afin de mettre en évidence le positionnement de la $n^{ème}$ région. Les régions représentées

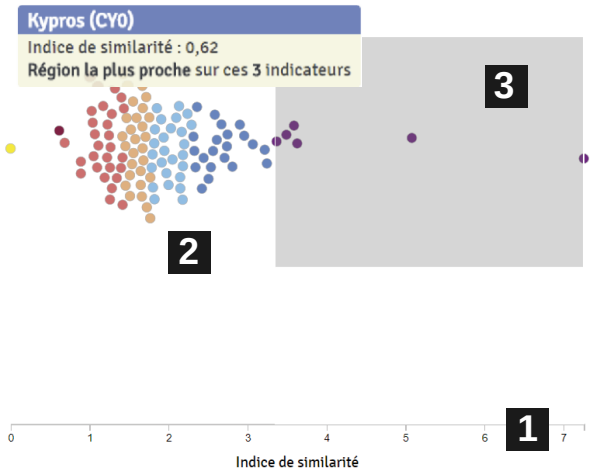
POSITION
1 ind. 2 ind. +3 ind.

RESSEMBLANCES
+1 ind.

QUELLES RÉGIONS ?

Type de similarité:
- Écart de rang moyen
- Distance euclidienne

1

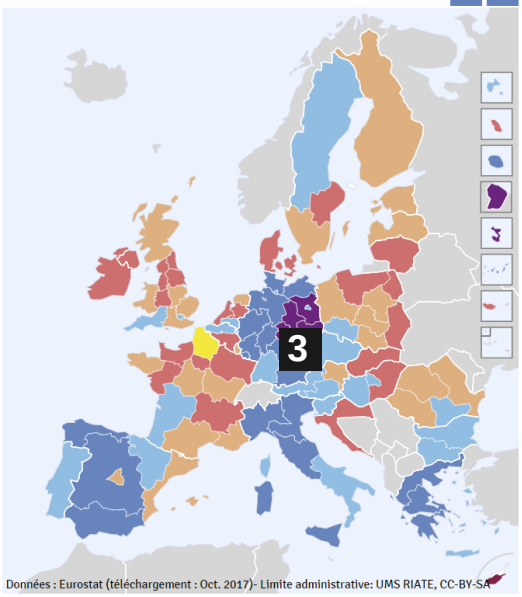


4 Ressemblance globale Par indicateur

Cercles proportionnels à la population

Ensemble des régions
Données disponibles pour 107/108 régions.

2

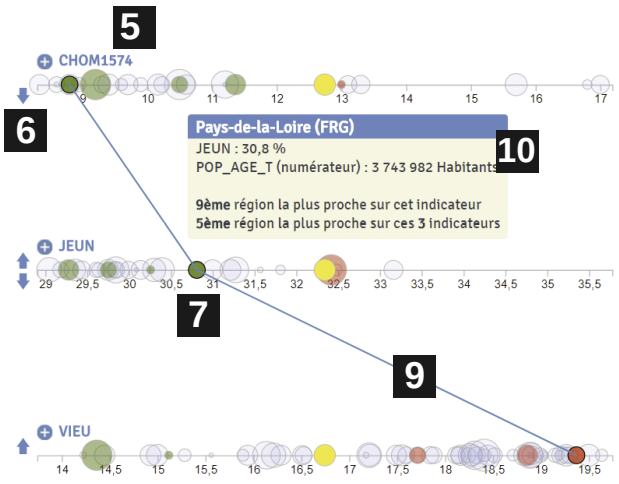


- Ma région : Hauts-de-France
- Région la plus ressemblante
- Régions très ressemblantes
- Régions ressemblantes
- Régions dissemblantes
- Régions très dissemblantes

2

Variable	Min.	Moy.	Med.	Max.	Ma région
CHOM1574	2,6	9,4	7,1	27,5	12,7
JEUN	19,3	27,3	27,1	60,6	32,3
VIEU	2,6	18,8	19	25,1	16,7

5



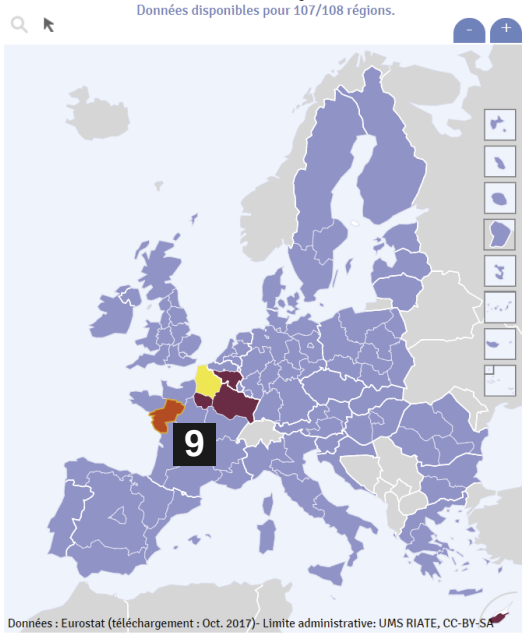
Ressemblance globale Par indicateur 4

Sélection de la 5 ème région la plus proche 8

7 Cercles proportionnels à la population

Classement par degré de ressemblance 11

Ensemble des régions
Données disponibles pour 107/108 régions.



- Ma région : Hauts-de-France
- Autres régions de l'espace d'étude (sélectionnables)

Variable	Min.	Moy.	Med.	Max.	Ma région
CHOM1574	2,6	9,4	7,1	27,5	12,7
JEUN	19,3	27,3	27,1	60,6	32,3
VIEU	2,6	18,8	19	25,1	16,7

en gris sont celles dont les valeurs sont proches de ma région pour un indicateur, mais très éloignées sur les autres indicateurs.

9 En cliquant sur un cercle ou une région, celle-ci apparait en surbrillance et un « éclair » apparait pour rappeler le positionnement de la région par rapport à ma région sur chacun des axes.

10 Le survol d'une des régions sur le graphique permet l'affichage d'une boite d'aide qui rappelle son niveau de ressemblance avec ma région pour l'ensemble des indicateurs et pour cet indicateur spécifiquement.

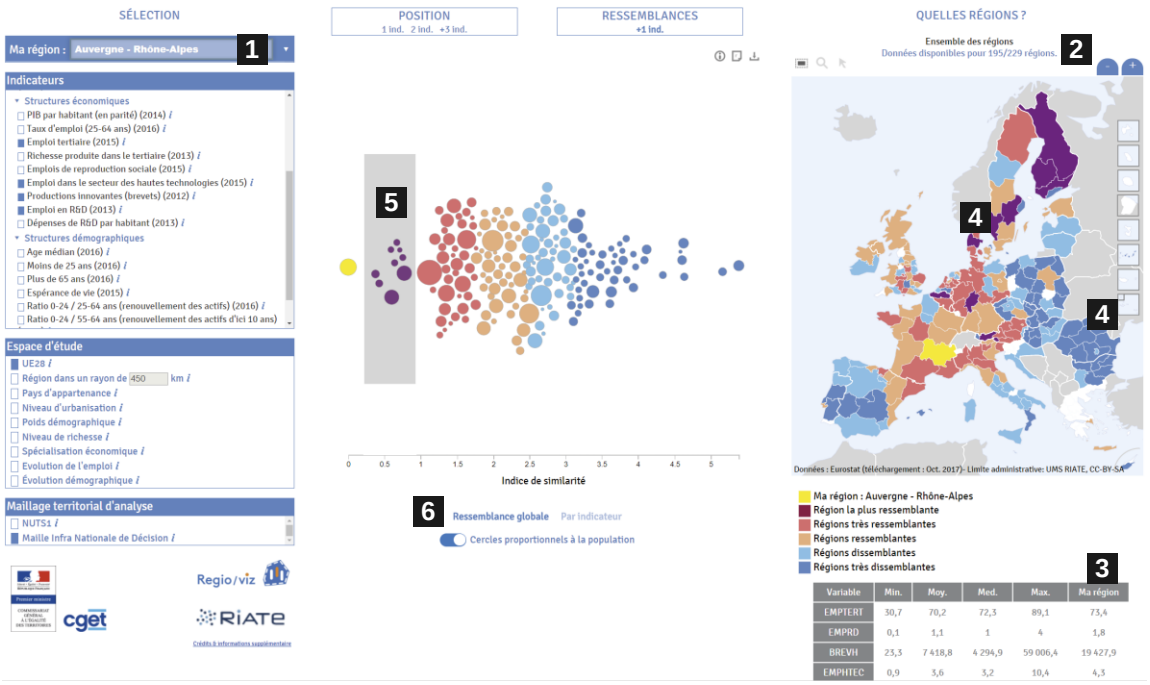
11 L'activation de l'option « classement par degré de ressemblance » permet de visualiser pour la X^{ème} région la plus proche les indicateurs dont les valeurs sont les plus ressemblantes de l'unité territoriale de référence (rang sur les écarts statistiques observés indicateur par indicateur). Une fois cette option activée, l'indicateur qui introduit le plus de ressemblance est replacé en haut de la fenêtre graphique, l'indicateur le moins ressemblant en bas.

Un exemple d'utilisation : Emploi et innovation en région Auvergne-Rhône-Alpes

L'objectif de ce scénario consiste à identifier les régions qui ressemblent le plus à la région Auvergne-Rhône-Alpes sur la thématique de l'innovation et de l'emploi à haute valeur ajoutée. Quatre indicateurs ont été identifiés pour procéder à cette comparaison : l'emploi tertiaire (2015), l'emploi dans le secteur des hautes technologies (2015), les productions innovantes (2012) et l'emploi en R&D (2013). L'analyse porte sur la maille infranationale de décision (MIND) qui comprend 229 unités territoriales en combinant NUTS1 et NUTS2 comme maille d'observation.

1 La première étape consiste à paramétrer Regioviz pour pouvoir réaliser l'analyse. La région Auvergne-Rhône-Alpes, les 4 indicateurs, l'espace d'étude UE28 et le maillage territorial d'analyse « MIND » sont retenus pour cette analyse. La méthode de mesure « distance euclidienne » est sélectionnée afin d'évaluer les distances statistiques réelles qui séparent Auvergne-Rhône-Alpes du reste des régions européennes.

2 L'analyse de complétude de l'information située au-dessus de la carte rappelle que compte-tenu des indicateurs sélectionnés, les données sont disponibles pour 195 des 229 régions de l'espace d'étude, soit 96,4 % de la population de l'espace d'étude (clic gauche sur cet élément textuel). En particulier, les données sont indisponibles pour les DROM et la Corse, la Slovénie, la plupart des régions grecques ainsi que quelques régions polonaises et britanniques (ces régions apparaissent en blanc sur la carte).



3 Ce tableau récapitulatif (qu'il convient aussi d'approfondir avec les autres modules d'exploration de Regioviz) rappelle que la région Auvergne-Rhône-Alpes se situe nettement au-dessus de la moyenne pour les 4 indicateurs de l'analyse : part de l'emploi tertiaire de 73,4 % (70,2 % pour la moyenne européenne de l'UE28 – hors données manquantes), part de l'emploi dans le secteur des hautes technologies de 4,3 % (3,6 % pour l'UE28), part de l'emploi en recherche et développement de 1,8% (1,1 % pour l'UE28) et brevets déposés par millions d'habitants de 19428 (7419 pour l'UE28).

4 Le graphique permet de visualiser les régions proches et éloignées d'Auvergne-Rhône-Alpes pour les indicateurs retenus. Les régions les plus ressemblantes (rouge foncé) se situent l'arc nord méditerranéen (Catalogne, Occitanie, PACA, Piémont, Ligurie), en Autriche, en Belgique, dans le nord-ouest de l'Allemagne et en Scandinavie. Les régions les moins ressemblantes sont situées en Europe Centrale et Orientale et dans le sud-ouest de la péninsule ibérique. La région la moins ressemblante est la région du nord-est roumain (indice de similarité de 5,39).

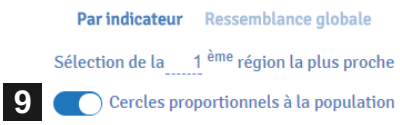
5 Le positionnement des régions sur l'axe de similarité permet de distinguer la distance statistique séparant Auvergne-Rhône-Alpes des autres régions européennes. Un saut de similarité apparaît entre la 9ème et la 10ème région, laissant présager que 9 régions ressemblent nettement plus à Auvergne-Rhône-Alpes sur ces indicateurs que les autres. La sélection sur le graphique permet d'identifier qu'il s'agit respectivement des régions de Västsverige (Suède, indice de similarité 0.38), Östra Mellansverige (Suède, 0.43), Etalä-Suomi (Finlande, 0.59), Hessen (Allemagne, 0.6), Länsi-Suomi (Finlande, 0.64), Midtjylland (Danemark, 0.67), Tirol (Autriche, 0.71), Pohjois-ja Itä-Suomi (Finlande, 0.76) et Vlaams Gewest (Belgique, 0.77). Les régions les plus proches d'Auvergne-Rhône-Alpes pour cette sélection correspondent donc majoritairement à des régions non-capitales de Scandinavie.

6 Cet indice n'en reste pas moins synthétique et peut cacher une importante hétérogénéité statistique lorsque l'on considère ces quatre indicateurs séparément. Pour en connaître davantage sur les facteurs qui font que ces régions sont proches statistiquement il convient d'utiliser la fonction « ressemblances par indicateur ».

7 Un nouveau module est activé et caractérise par défaut le profil de deux régions : la région Auvergne-Rhône-Alpes et la région suédoise de Västsverige qui lui ressemble le plus d'après l'indice global de similarité. Une première étape consiste ici à améliorer la lisibilité du graphique. Les couleurs sont inversées de telle sorte que la couleur du rond rouge signifie que l'autre région soit dans une situation plus défavorable sur l'indicateur

8 L'ordre d'apparition des axes est aussi modifié pour que tous les indicateurs pour lesquels la région Auvergne-Rhône-Alpes se positionne en situation favorable se succèdent sur le graphique. La région Auvergne-Rhône-Alpes se trouve dans une situation favorable par rapport à la région suédoise sur la part de l'emploi dans les hautes technologies et sur le nombre de brevets déposés par million d'habitants. Elle se situe dans une position défavorable par la part de l'emploi dans le tertiaire et dans les secteurs de la recherche et du développement.

9 L'option « cercles proportionnels à la population » permet d'avoir une idée de la différence de poids qui existe entre ces deux régions. Toute chose égale par rapport au temps d'emploi en R&D, la région Auvergne-Rhône-Alpes est ainsi caractérisée par un nombre d'emplois dans le secteur de la recherche et développement bien plus important (58 000) que son homologue suédoise (17 160).



10 En plus des informations relatives au poids de chacune des régions contenues dans la boite d'aide, celle-ci révèle également que ces deux régions sont très ressemblantes pour trois des quatre indicateurs : la région Västsverige est la 3ème région la plus proche d'Auvergne-Rhône-Alpes pour la part de l'emploi en recherche et développement et le nombre de brevets par million d'habitants. C'est la 6ème pour la part de l'emploi dans le secteur des hautes technologies. En revanche ces deux régions sont statistiquement éloignées pour la part d'emploi tertiaire : c'est la 64ème région la plus proche (73.4 % pour Auvergne-Rhône-Alpes contre 77.1 % pour la région suédoise).