JOURNAL DE LA SOCIÉTÉ STATISTIQUE DE PARIS

J. DUFRÉNOY

La relation entre la distance spatiale et la distance lexicale

Journal de la société statistique de Paris, tome 113 (1972), p. 146-150

<http://www.numdam.org/item?id=JSFS_1972__113__146_0>

III

VARIÉTÉS

LA RELATION ENTRE LA DISTANCE SPATIALE ET LA DISTANCE LEXICALE (1)

M. J. Seguy imagine, à l'époque des dialectes locaux (p. 352), un voyageur partant de son lieu d'origine où se parle un certain dialecte défini par un atlas linguistique; à mesure qu'il s'éloigne de distances spatiales x_i km croissantes, il rencontre de plus en plus de différences entre le dialecte au lieu d'origine ($x = 0$) et le dialecte à x_i km.

Avec la coopération de M. Henri Guitier, professeur de linguistique romane à l'Université de Montpellier, M. J. Seguy a comparé divers modèles mathématiques permettant de représenter la variation de la variable dépendante y, représentant la « distance lexicale » exprimée en « distance de Hamming » (selon le système binaire de tout ou rien) en fonction de la variation de la variable indépendante x km; y constitue l'indice de l'intercompréhension; jusque vers $y = 50$ la compréhension fonctionne encore (p. 344); tant que le dialecte est fonctionnel, c'est-à-dire sert aux communités en contact à communiquer tout en se démarquant, la distance lexicale se maintient au-dessous de 50 (p. 351).

1° *Modèle de fonction exponentielle*

J. Seguy considère le gallo-roman comme un ensemble dont les éléments sont les dialectes, ensemble où agit un facteur d'*auto*-régulation matérialisé par la courbe : $y = K \log (x + 1)$ laquelle est un avatar de la loi d'économie toujours présente dans les phénomènes linguistiques.

Pour celui qui parcourt un itinéraire on peut supposer que pour toute distance parcourue dx la probabilité de manifestation de difficulté linguistique est proportionnelle à la longueur de l'intervalle et la distribution des probabilités est conforme à celle d'une fonction exponentielle pouvant se représenter par une courbe tendant asymptotiquement vers un plafond correspondant au niveau $y \sim 55$ à partir duquel l'intercompréhension cesse de fonctionner.

Exemple numérique : Tableau ALF_I : Itinéraire Roussillon-Belgique; à chaque distance kilométrique x de 24 à 400 km, on fait correspondre la valeur y de la distance lexicale; on cumule ces valeurs de y de haut en bas pour obtenir les y cumulés; on transforme les y cumulés en pourcentages cumulés n en fonction d'un total de 500, pris pour 100 %. On porte chaque pourcentage n cumulé, à son niveau d'ordonnées, sur l'échelle doublement logarithmique du papier de Weibull; et à son rang d'abscisses sur échelle des log x.

On définit, pour x compris entre 9 et 356, neuf points par lesquels on peut faire passer une droite de régression de pente 1,15, très voisine de 1, pente de la droite qui sur papier de Weibull représente l'anamorphose en droite d'une courbe exponentielle.

1. Article de M. J. Séguy paru dans la revue de linguistique romane 139-140, juillet-décembre 1971, 334-357.

TABLEAU ALF_I

x	y	y cumulés	n (% cumulés)
0	0	0	0
24	9	9	1,8
51	37	46	9
86	40	86	17
135	40	126	25
182	47	173	35
246	45	218	44
293	52	270	54
318	52	322	64
356	56	378	76
384	62	430	86
415	57	487	98
450	55	542	

2° *Phénomènes de récurrence*

L'univers que parcourt l'individu partant de son lieu d'origine vers des lieux de plus en plus éloignés peut être comparé à l'ensemble de 2 R boules numérotées successivement de 1 à 2 R, distribuées dans deux urnes I et II. On tire au hasard un nombre compris entre 1 et 2 R et on tire la boule portant le numéro correspondant de l'urne où elle se trouve pour la transférer dans l'autre urne; on effectue successivement S tirages indépendants, dont les résultats successifs peuvent se représenter par une courbe dont la première partie, exponentielle, tend vers le niveau d'égalisation, pour ensuite manifester des variations quasi périodiques, conformément au phénomène de récurrence de Poincaré.

Telle est d'ailleurs l'allure de la courbe $y = f(x)$ représentant les résultats des différents parcours.

3° *Modèles de régression parabolique*

Comparaison de 2 itinéraires, l'un défini par ALF, Atlas linguistique de France, par Gilliéron et l'autre par ALMC, Atlas linguistique et ethnographique du Massif Central, par P. Nauton (pp. 354, 355).

L'auteur indique les valeurs de y pour des distances x km formant approximativement les termes de la progression arithmétique 5, 10, 15, 20; ce qui permet de calculer rapidement l'équation polynomiale sous la forme

$$\hat{y} = \bar{y} + BZ_1 + CZ_2 \text{ où } Z_1 = 2(x - \bar{x}) \quad \text{et} \quad Z_2 = \left[(x - \bar{x})^2 - \frac{n^2 - 1}{12}\right]$$

avec

$$n = 4;\ B = \frac{\Sigma (Z_1 y)}{\Sigma (Z_1)^2} \quad \text{et} \quad C = \frac{\Sigma (Z_2 y)}{\Sigma (Z_2)^2}$$

x	Z_1	Z_2	y_{ALF}	Z_1y	Z_2y	$\hat{y}$	y_{ALMC}	Z_1y	Z_2y	$\hat{y}$
5	—3	1	15	—45	15	14,85	27	—81	27	27,5
10	—1	—1	20	—20	—10	19,95	44	—44	—44	42,5
15	1	—1	25	25	—25	25,05	50	50	—50	51,5
20	3	1	31	91	31	30,65	55	165	55	54,5
			91	51	—1		176	90	12	176

La première équation $\hat{y}_{ALF} = 22{,}75 + 2{,}55\ Z_1 - 0{,}25\ Z_2$ confirme ce que suggérait le simple examen du tableau :

y_{ALF} croît presque linéairement en fonction de x, dans l'étendue $x = 5$ à $x = 20$.

La seconde équation $\hat{y}_{ALMC} = 44 + 4{,}5\ Z_1 - 3\ Z_2$ suggère une augmentation de y_{ALMC} moins que proportionnelle à l'augmentation de x, le coefficient quadratique (—3) étant numériquement voisin du coefficient linéaire (4,5).

On peut d'ailleurs introduire cette contrainte : « la courbe doit passer par l'origine avec $y = 0$ pour $x = 0$, on calcule alors l'équation polynomiale $\hat{y} = 31{,}5 + 3{,}27\ Z_1 - 1{,}2\ Z_2$ selon le tableau suivant pour les valeurs de x formant les termes de la série 0, 1, 2, 4; correspondant aux valeurs de x.

x	y	Z_1	Z_1y	Z_2	Z_2y	$\hat{y}$
0	0	—7	0	7	0	0,28
5	27	—3	—81	—4	—108	26,49
10	44	1	44	—8	—352	44,37
20	55	9	95	5	275	54,98
	126		458		—185	126,32

$$\bar{y} = \frac{126}{4} = 31{,}5;\ B = \frac{\Sigma\,(Z_1y)}{\Sigma\,(Z_1)^2} = \frac{458}{140};\ C = \frac{\Sigma\,(Z_2y)}{\Sigma\,(Z_2)^2} = -\frac{185}{154}$$

Analyse de la variance

Source de variation	Somme des carrés	Degrés de liberté	
Totale $\Sigma\ (yi - \bar{y})^2$	1 720	4	
Due à régression linéaire $B\ \Sigma\ (Z_1y)$ . . . = (3,27) (458)	1 497,6	1	$F = \frac{1\,497{,}6}{0{,}34} = 443;\ p \sim 0{,}96$
Quadratique $C\ \Sigma\ (Z_2y)$ = (—1,2) (—185)	222,0	1	
Résidu $\Sigma\ (y - \bar{y})^2$	0,34	1	

La contribution de la régression linéaire est très significative avec $F = 443$ pour 1 degré de liberté au numérateur, 1 degré de liberté au dénominateur, mais celle de la régression quadratique n'est pas significative.

Pour faire apparaître la signification de la contribution du terme quadratique nous utilisons la série suivante des x

x	Z_1	Z_2	y	Z_1y	Z_2y	$\hat{y}$
0	—2	2	0	0	0	0,74
5	—1	—1	27	—27	—27	25,83
10	0	—2	44	0	—88	43,36
15	1	—1	50	50	—50	52,43
20	2	2	55	110	110	53,94
			176	133	55	

$$\hat{y} = \frac{176}{5} = 35{,}2;\ B = \frac{\Sigma\,(Z_1y)}{\Sigma\,(Z_1)^2} = \frac{133}{10};\ C = \frac{\Sigma\,(Z_2y)}{\Sigma\,(Z_2)^2} = \frac{55}{14}$$

$$\hat{y} = 35{,}2 + 13{,}3\ Z_1 - 3{,}93\ Z_2$$

Analyse de la variance

Source de variation	Somme des carrés	Degrés de liberté	Carré moyen	
Totale $\Sigma\ (y-\bar{y})^2$	2 000	4		
Due à régression linéaire $B\Sigma\ (Z_1\ y)$	1 769	1	1 769	$F = \frac{216}{7,5} \sim 30;\ p \sim 0,97$
Quadratique $C\Sigma\ (Z_2\ y)$	216	1	216	
Résidu.	15	2	7,5	

Pour 16 parcours étudiés, J. Seguy trouve une relation fonctionnelle $y = f(x)$ ayant la forme d'une fonction logarithmique : par exemple, pour l'itinéraire « Pouilles-Alpes », en utilisant Sprach — Und Sach Atlas Italien und der Südschweiz de Jaberg et Jung (AIS), J. Seguy calcule $y = 25 \log (x + 1)$.

Dans AIS, selon M. Guitier, les croissances se font par paliers, chaque saut correspondant à la frontière d'un domaine dialectal italien.

Quant aux « dimensions spatiales », le parcours AIS 1, partant du sud des Pouilles pour remonter vers le nord jusqu'aux Alpes valdotaines, couvre 1 100 km.

Emploi de l'échelle arc.tangente

Si nous transformons les distances spatiales x km en $x' = \frac{2}{\pi} M \arctan [x/K\eta]$ nous remplaçons l'étendue des distances spatiales $x = 0$ à $x = 1\,100$, par l'étendue des variations de $x' = 0$ à $x' = M$ sur l'échelle arc tan x, telle que la distance 0 à $M/2$ mesurée sur cette échelle arc tan x, soit la même que la distance M/2 à M, elle aussi mesurée sur l'échelle arc tan x; le graphique montre que $M/2$ se situe à $x = 250$; De chaque valeur x'_i de la transformée, nous élevons une verticale de hauteur y_i; par les points (x'_i, y_2) nous traçons une droite de régression qui intercepte le niveau $y_{n/2}$ au niveau $32,5 = \frac{(46\text{-}23)}{2}$; on rencontre donc 50 % de la distance lexicale dans le parcours $x' = 0$ à $x' = 7/2$ c'est-à-dire $x = 0$ à $x = 250$ km.

Conclusions

1° La relation la plus vraisemblable entre distance lexicale y et distance géographique x peut se représenter par une courbe exponentielle, qui se prête à anamorphose en droite de pente 1 sur papier-loi de Weibull.

2° Pour des trajets de moins de 20 km, le tableau, page 355, permet de calculer la relation fonctionnelle entre x et y selon une équation polynomiale de second degré en x; pour d'aussi faibles variations d'étendue il est difficile de choisir le modèle le plus vraisemblable « courbe exponentielle » ou « arc de parabole ».

3° Au lieu de rechercher le modèle le plus vraisemblable permettant de représenter la fonction $y = f(x)$ dans l'étendue $x = 0$ à $x = 500$ ou 1 000 km, on peut considérer qu'après une phase initiale où la fonction peut se représenter par une courbe exponentielle ou par un arc de parabole, intervient une phase durant laquelle y tend vers un niveau de saturation avec paliers et oscillations. Au cours de cette phase, la transformation de la

variable dépendante x en arc tan x permet d'obtenir un alignement linéaire des points (arc tan x, y).

L'étude effective par M. Seguy, quant à la relation entre la distance spatiale et la distance lexicale est d'autant plus opportune que se complète la collection des atlas linguistiques de la France par régions notamment par l'atlas du Centre de M^lle^ Dubuisson, et l'atlas de l'Ouest de M^lle^ Massignon et M^lle^ Horiot, et que paraissent des ouvrages tel que le vocabulaire de géographie agraire, par P. Fénelon.

J. Dufrénoy