

ANALYSE DES VARIATIONS REGIONALES DU DIVORCE EN TCHECOSLOVAQUIE EN 1980

Christina ASCHAN-LEYGONIE, Université Paris I
Claude GRASLAND, CNRS-Université Paris I

Type : EVALUATION

Niveau : Fin de première année de DEUG

Durée : 4 heures.

Thèmes : STATISTIQUE, CARTOGRAPHIE, DEMOGRAPHIE

Objectifs :

Les enseignements d'initiation à la statistique et à la cartographie sont complémentaires. L'Université Paris 1 les regroupe à l'intérieur d'une même unité de valeur en première année de DEUG et les enseignants des deux disciplines travaillent en étroite collaboration. L'évaluation des connaissances se fait de façon séparée lors du premier partiel (Février) mais de façon conjointe lors du second (Juin). C'est cette évaluation commune que nous présentons ici, le sujet ayant été proposé à 130 étudiants de première année de DEUG de Géographie en Juin 1994.

L'objectif de cette évaluation était, d'une part, de vérifier la capacité des étudiants à appliquer les outils statistiques et cartographiques à un problème géographique, d'autre part, de s'assurer qu'ils ont fait le lien entre les deux types d'enseignement. On a volontairement réduit la part des calculs pour privilégier la qualité de l'analyse géographique d'un problème à l'aide de différents outils. Le travail se présente donc comme une micro-étude devant déboucher sur des résultats concrets (connaissance de la répartition et des mécanismes du divorce en Tchécoslovaquie en 1980).

Nous présentons successivement le sujet proposé (pages 2 à 4), le corrigé (pages 5 à 9) et les résultats obtenus face aux étudiants (page 10).

Analyse des variations régionales du divorce en Tchécoslovaquie en 1980

Sujet

Durée : 4 heures

Barème : 40 pts (20 pts en statistique + 20 pts en cartographie)

1°) Choix de l'indicateur

Pour analyser les variations régionales du divorce en Tchécoslovaquie en 1980, on a retenu le nombre de divorces pour 1000 mariages (noté TDIVO dans le tableau de données).

a) Précisez le type et la nature de ce caractère.

b) Pourquoi est-il plus intéressant que le nombre absolu de divorces (DIVO) ou le nombre de divorces par habitant (DIVO/POPUL) ?

2°) Analyse de la distribution des divorces et de l'urbanisation

On se propose de comparer les distributions spatiales du taux de divorce pour 1000 mariages (TDIVO) et du taux d'urbanisation (TURBA), afin d'examiner l'hypothèse d'une relation entre ces deux variables.

a) Calculez la médiane de ces deux caractères. Dans chaque cas, que vous apprend la comparaison de la moyenne et de la médiane sur la forme de la distribution ?

b) Déduisez-en deux partitions en quatre classes permettant de cartographier ces deux caractères et de les comparer. Justifiez le choix des classes retenues.

c) Cartographiez les deux caractères en noir et blanc.

N.B. Les cartes sont à réaliser à l'encre sur calque. Les trames choisies ne devront pas utiliser les à-plats noir et blanc.

3°) Analyse de la relation divorce-urbanisation

a) Commentez les deux cartes que vous venez de réaliser. Que remarquez-vous ?

b) Si l'on fait l'hypothèse que le taux de divorce dépend du taux d'urbanisation, quelle sera l'équation de la droite de régression exprimant leur relation moyenne (N.B. le coefficient de corrélation linéaire des variables TURBA et TDIVO est égal à +0.82) ?

c) Tracez cette droite sur le graphique cartésien des deux caractères.

4°) Analyse des écarts à la relation divorce-urbanisation

a) Calculez algébriquement ou graphiquement les résidus des régions Bohême-Ouest (BO) et Slovaquie-Centre (SC) puis complétez le tableau B.

b) Commentez les résidus de régression des régions de Bohême-Centre (BC), Bohême-Sud (BS) et Bratislava (BR). Que signifient-ils ?

c) Plus généralement, comment pouvez-vous interpréter le fait que toutes les régions slovaques ont des résidus négatifs et toutes les régions tchèques (sauf une) des résidus positifs ?

DIVORCE ET URBANISATION EN TCHECOSLOVAQUIE EN 1980

A) CODE DES VARIABLES

NOM : nom de la région
 CODE : code cartographique
 REP : T = république tchèque ; S = république slovaque
 POPUL : population totale en 1980 (en hab.)
 PURBA : population urbaine en 1980 (en hab.)
 DIVO : nombre de divorces en 1980
 MARI : nombre de mariages en 1980
 TURBA : taux de population urbaine en 1980 (en %)
 TDIVO : divorces pour 1000 mariages en 1980
 RESID : résidu de la régression TDIVO = a * TURBA + b

B) TABLEAU DES DONNEES

Source : recensement de population de 1980

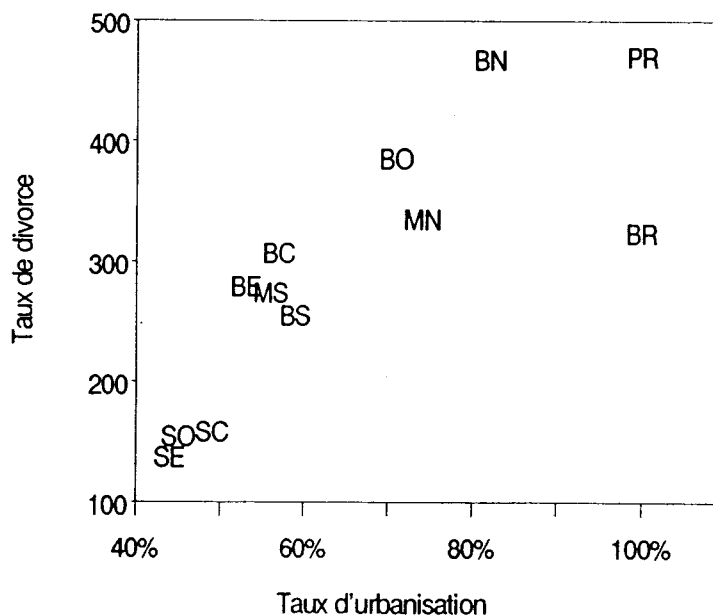
NOM	CODE	REP	POPUL	PURBA	DIVO	MARI	TURBA	TDIVO	RESID
PRAGUE	PR	T	1190206	1190206	4239	8956	100 %	473	18
BOHEME-CENTRE	BC	T	1148262	653361	2579	8306	57 %	310	54
BOHEME-SUD	BS	T	690015	407109	1294	5022	59 %	258	-8
BOHEME-OUEST	BO	T	888152	627035	2847	7342	71 %	388
BOHEME-NORD	BN	T	1172563	961502	4426	9418	82 %	470	98
BOHEME-EST	BE	T	1251461	667029	2579	9138	53 %	282	44
MORAVIE-SUD	MS	T	2041761	1143386	4232	15305	56 %	277	25
MORAVIE-NORD	MN	T	1944372	1434947	5022	14856	74 %	338	3
BRATISLAVA	BR	S	379437	379437	1094	3350	100 %	327	-128
SLOVAQUIE-OUEST	SO	S	1685393	756741	2102	13380	45 %	157	-44
SLOVAQUIE-CENTRE	SC	S	1526363	749444	1909	11822	49 %	161
SLOVAQUIE-EST	SE	S	1393136	614373	1540	11026	44 %	140	-56
Moyenne (ensemble)	***	***	1275927	798714	2822	9827	66 %	298	0
Ecart-type (ensemble)	***	***	466344	308034	1286	3515	19 %	107	61
Moyenne (T)	***	***	1290849	885572	3402	9793	69 %	350	38
Ecart-type (T)	***	***	441607	328690	1182	3322	15 %	80	33
Moyenne (S)	***	***	1246082	624999	1661	9895	60 %	196	-72
Ecart-type (S)	***	***	510943	152689	385	3872	23 %	76	33

C) GRAPHIQUE CARTESIEN DE TURBA ET TDIVO

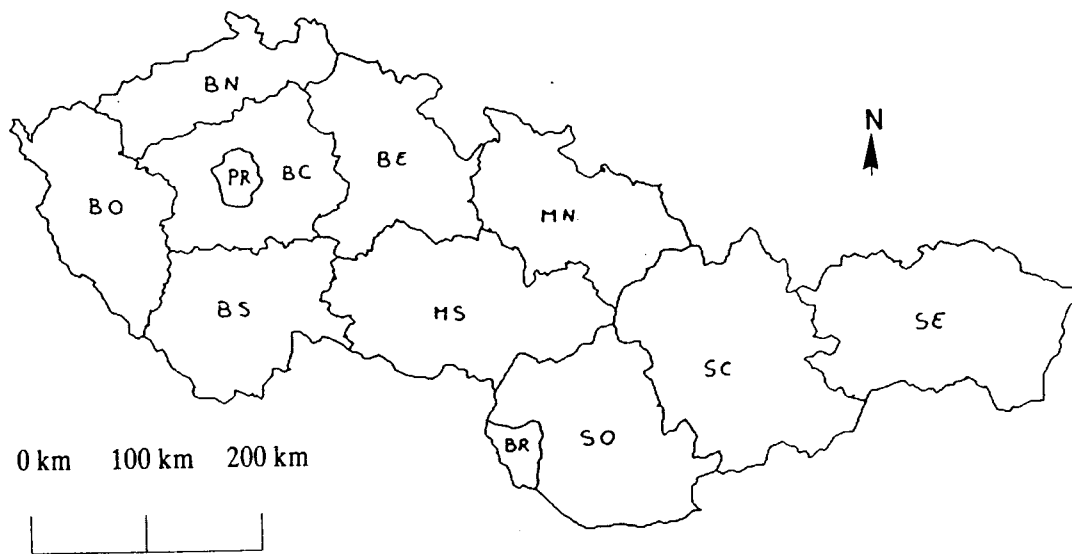
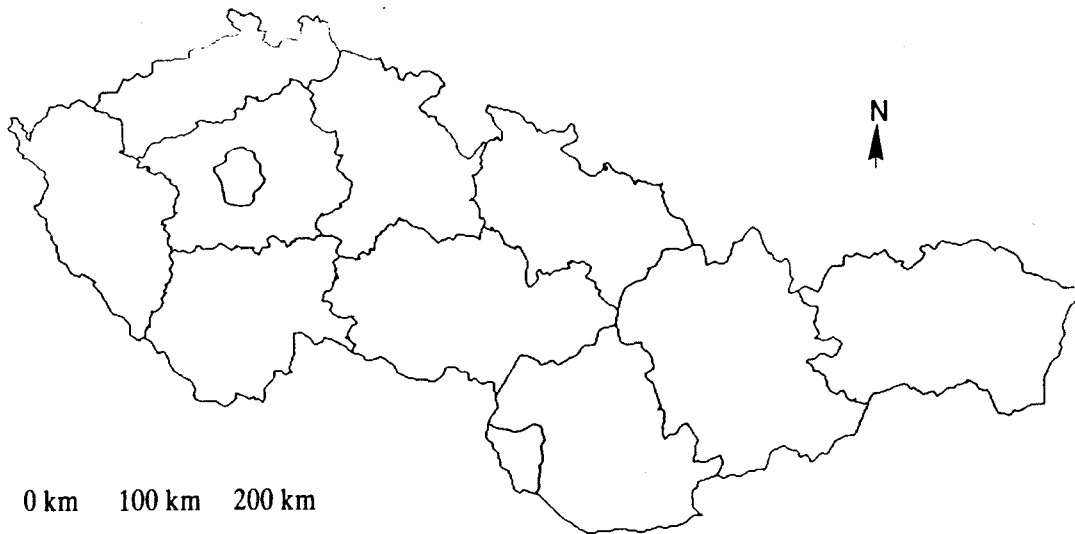
Le coefficient de corrélation linéaire des variables TURBA et TDIVO est :

$r = +0.82$

LES REGIONS TCHECOSLOVAQUES EN 1980



LES 12 REGIONS TCHECOSLOVAQUES (KRAJS) EN 1980



REPUBLIQUE TCHEQUE		SLOVAQUIE (REP.)
PR : Prague	BO : Bohême-Ouest	BR : Bratislava
BC : Bohême-Centre	BE : Bohême-Est	SO : Slovaquie-Ouest
BN : Bohême-Nord	MS : Moravie-Sud	SC : Slovaquie-Centre
BS : Bohême-Sud	MN : Moravie-Nord	SE : Slovaquie-Est

Analyse des variations régionales du divorce en Tchécoslovaquie en 1980

Barème et Correction

* STATISTIQUES (20 pts)

1°) a) Le caractère TDIVO est **quantitatif** (la moyenne a un sens), **mesurable** (la valeur 0 n'est pas arbitraire), **continu** (il y a une infinité de valeurs sur un intervalle) et il s'agit d'un **taux** (la somme des modalités n'a pas de signification).

=> 2pts : soit 0.5 pts par attribut correct

1°) b) Le nombre absolu de divorces n'est pas intéressant car il subit un **effet de taille** (1000 divorces n'ont pas le même sens dans une région de 100 000 hab. ou une région de 1 000 000 hab.).

- Le rapport du nombre de divorces sur le nombre d'habitants est également peu intéressant car il inclut des **effets de structure** : en effet, on ne peut divorcer que si l'on est marié, de sorte que les régions ayant une forte proportion de non-mariés (jeunes, veufs, ...) risquent d'avoir des taux de divorces sous-estimés.

- Le taux retenu est donc la meilleure solution compte tenu de l'information disponible mais il présente un certain nombre de défauts. Il s'agit en effet d'un **indice conjoncturel** puisqu'il ne tient compte que des divorces et des mariages de l'année en cours. Un indice plus intéressant serait le rapport du nombre de divorces sur le nombre total de mariages fourni par le recensement, voire la probabilité de divorce par années écoulées depuis le mariage.

=> 2 pts : soit 1 pt par argument correct (le 3e argument est facultatif puisqu'il n'est pas posé explicitement dans le sujet).

2°) a) La distribution comportant un nombre pair d'éléments (12), la médiane correspondra au centre de l'intervalle médian, c'est à dire à la moyenne des 6e et 7e valeurs classées par ordre d'importance. Soit :

TURBA : intervalle médian [57 ; 59] et **médiane = 58 %**

TDIVO : intervalle médian [282 ; 310] et **médiane = 296 p. 1000**

- Les moyennes étant fournies dans le tableau B, il est inutile de les calculer. Dans chaque cas, la comparaison de la moyenne et de la médiane permet de savoir si la distribution est symétrique (les deux valeurs centrales sont proches) ou dissymétrique (les deux valeurs centrales sont éloignées). L'éloignement des deux valeurs centrales doit être jugé par référence à la dispersion des distributions (étendue ou écart-type). Soit :

TURBA : moyenne > médiane. La différence est de 8 pour un écart-type de 19 et une étendue de 66. Elle est donc assez importante et on peut affirmer que **la distribution est nettement dissymétrique à gauche**, c'est-à-dire qu'il y a **concentration dans les valeurs faibles et dispersion dans les valeurs fortes**. Cette dissymétrie résulte des deux valeurs exceptionnelles des régions-villes de Prague et Bratislava (100%).

TDIVO : moyenne \approx médiane. La différence n'est que de 2 pour un écart-type de 107 et une étendue de 333. Les deux valeurs sont donc pratiquement confondues et on peut affirmer que **la distribution est symétrique**, c'est-à-dire que **les valeurs se distribuent équitablement de part et d'autre de la moyenne et de la médiane**.

=> 4 pts, soit 2 pts pour le calcul des médianes et 2 pts pour la détermination de la forme des distributions. A noter que beaucoup d'étudiants ont été sanctionnés pour avoir affirmé que les distributions étaient unimodales alors que la simple prise en compte de la moyenne et de la médiane ne permet pas de le savoir.

2°) b) Lorsque l'on cherche à comparer deux distributions, il est nécessaire d'utiliser un principe identique de découpage des classes.

- **La méthode des "seuils naturels"** est donc à exclure puisqu'elle respecte l'identité de chaque distribution et ne suit pas un principe commun. De plus, le choix du nombre de classes étant imposé (4), il est peu probable de trouver précisément trois seuils dans chaque distribution.

- **La méthode des amplitudes égales** (basée sur la division de l'étendue en quatre classes de même amplitude) n'est valable que lorsque les distributions sont homogènes et ne comportent pas de valeurs exceptionnelles. Cette méthode est par ailleurs assez fragile puisqu'elle ne dépend que du minimum et du maximum. Elle ne sera donc pas retenue, d'autant que l'analyse précédente a montré la présence de valeurs exceptionnelles dans la distribution de TURBA.

- **La méthode de la moyenne et de l'écart-type** (moyenne comme limite de classe et écart-type comme amplitude de classe) a l'avantage de permettre une comparaison standard et possède une signification probabiliste (66.7 % des régions seront comprises dans les deux classes centrales si les distributions sont normales - gaussiennes -). Elle est toutefois à déconseiller lorsque les distributions sont dissymétriques ou possèdent des valeurs exceptionnelles susceptibles d'influer sur le calcul de la moyenne et de l'écart-type (TURBA).

- **La méthode des effectifs égaux (quantiles)** est la plus robuste puisqu'elle ne tient compte que de l'ordre des valeurs. Elle est particulièrement conseillée lorsque les distributions sont de formes différentes, comme c'est le cas ici. On déterminera donc les quartiles : Q1, Q2 (Médiane) et Q3 qui partagent la distribution en quatre classes d'effectifs égaux. Le fait que la distribution ait un nombre d'éléments multiple de 4 facilite d'ailleurs le calcul (3 éléments par classe). On retiendra donc les partitions suivantes :

TURBA : [44 ; 51[, [51 ; 58[, [58 ; 78[, [78 ; 100]

TDIVO : [140 ; 210[, [210 ; 296[; [296 ; 363[; [363 ; 473]

=> 2 pts : à moduler en fonction de la justesse du choix retenu et des justifications apportées à ce choix.

3°) a) Les deux cartes présentent des configurations géographiques similaires, ce qui conduit à formuler l'hypothèse d'une corrélation positive, confortée par l'analyse du coefficient de Pearson ($r = + 0.82$). Pour chacune des deux variables, on observe globalement un gradient général de décroissance du NO vers le SE. Ce gradient n'est toutefois pas parfaitement régulier en raison de la présence de régions urbaines ou industrielles qui ont des taux de divorce ou d'urbanisation plus élevés que les régions voisines (Prague, Bratislava, Bohême-Nord). Au total, la distribution des deux variables recoupe assez bien la division du territoire en deux républiques : forte urbanisation et fort taux de divorce en République tchèque, faible urbanisation et faible taux de divorce en Slovaquie.

=> 2 pts , avec par exemple 1 pt pour la précision de la description géographique et 1 pt pour la formulation de l'hypothèse d'une relation divorce-urbanisation.

3°) b) Compte tenu de l'hypothèse retenue, le taux de divorce sera la variable dépendante (Y) et le taux d'urbanisation la variable indépendante (X). On va donc chercher l'équation de la droite de meilleur ajustement ($Y=a.X+b$) en supposant que le taux de divorces est une fonction linéaire du taux d'urbanisation. On sait que :

$$a = \text{Cov}(X,Y) / (\sigma_X)^2 \quad \text{et} \quad b = \bar{Y} - a\bar{X}$$

Comme les moyennes (\bar{X} et \bar{Y}) et les écart-types (σ_X et σ_Y) sont connus, la principale difficulté réside donc dans le calcul de la covariance qui est relativement long. Toutefois (le sujet prend bien soin de le préciser ...) on connaît le coefficient de corrélation des deux caractères $r(X,Y)$, ce qui permet aux étudiants astucieux d'éviter le calcul de la covariance. En effet :

$$r(X,Y) = \text{Cov}(X,Y) / (\sigma_X \cdot \sigma_Y) \Leftrightarrow \text{Cov}(X,Y) = r(X,Y) \cdot \sigma_X \cdot \sigma_Y \text{ d'où}$$

$$a = r(X,Y) \cdot \sigma_Y / \sigma_X = (0.82 \cdot 107) / 19 = 4.62$$

$$b = 298 - (4.62 \cdot 66) = -6.78$$

L'équation de la droite de régression est donc :

$$\text{TDIVO} = 4.62 \text{ TURBA} - 6.78$$

=> 2 pts si les calculs sont justes, quelle que soit la méthode employée (les étudiants qui ont calculé la covariance se sont sanctionnés eux-mêmes en perdant beaucoup de temps ...)

3°) c) Pour tracer la droite de régression il faut connaître deux points. On sait déjà (propriété de la droite de régression) qu'elle passe par le point moyen (66, 298). On calcule ensuite la valeur estimée de TDIVO pour une valeur quelconque de TURBA, ce qui donne par exemple les points (40, 178) et (100, 455). Il ne reste plus qu'à tracer la droite, en faisant toutefois attention à s'arrêter en (100, 455) car un taux d'urbanisation ne peut pas dépasser 100 % !

=> 1 pt si la droite est juste, avec une pénalité de 0.5 pts si le tracé a été poursuivi au-delà de 100 % d'urbanisation.

4°) a) Sachant que le résidu $Y_i - Y'_i$ est la différence entre la valeur observée Y_i et la valeur estimée $Y'_i = a.X_i + b$, le résultat est obtenu sans difficulté :

CODE	X	Y	Y'	Y-Y'
BO	71 %	388	321	+ 67
SC	49 %	161	220	- 59

=> 2 pts, soit 1 pt par calcul juste. Une pénalité de 0.5 pts a toutefois été appliquée lorsque les résultats comportaient un ou deux chiffres après la virgule (pseudo-précision vue l'information initiale).

4°) b) En Bohême-Centre, le résidu est positif, ce qui signifie que la droite de régression sous-estime le taux de divorces. En d'autres termes, le taux de divorces de cette région est plus élevé que ce que laisserait présager son taux d'urbanisation.

- A Bratislava, le résidu est négatif, ce qui signifie que la droite de régression sur-estime le taux de divorces. Le taux de divorce de cette région est donc moins élevé que ce que laisserait prévoir son taux d'urbanisation.

- En Bohême-Sud, le résidu est proche de 0, ce qui signifie que la droite de régression donne une estimation correcte du taux de divorces. Le taux de divorces est cette fois-ci conforme au taux estimé à partir du niveau d'urbanisation.

=> 2 pts, soit 1 pt pour le commentaire proprement statistique (sur ou sous-estimation) et 1 pt pour son interprétation empirique (niveau de divorce plus - ou moins - élevé que ce que laisserait présager le taux d'urbanisation).

4°) c) La régression linéaire a permis de dégager ce qui dans le phénomène divorce peut être expliqué par le niveau d'urbanisation. Le pouvoir explicatif de ce facteur est important ($r^2 = 67.2\%$) mais il subsiste des résidus qui indiquent le jeu possible d'autres facteurs. Le fait que les résidus ne soient pas disposés au hasard dans l'espace laisse penser que ces autres facteurs sont liés à l'appartenance à une république. Par exemple, on peut invoquer le fait que le développement de la Slovaquie a été plus tardif et qu'elle a moins été exposée aux influences occidentales que la République tchèque. On peut aussi rappeler que, bien que les deux républiques soient majoritairement catholiques, la proportion de croyants demeure nettement plus élevée en Slovaquie.

Au total, on peut donc conclure que le taux de divorces en Tchécoslovaquie dépend d'une part d'un facteur global (l'urbanisation) et d'autre part de facteurs historiques et culturels propres à chacune des deux républiques.

=> 1 pt si les hypothèses proposées sont acceptables et 1 pt de bonus si l'étudiant montre une bonne culture géographique sur le sujet étudié.

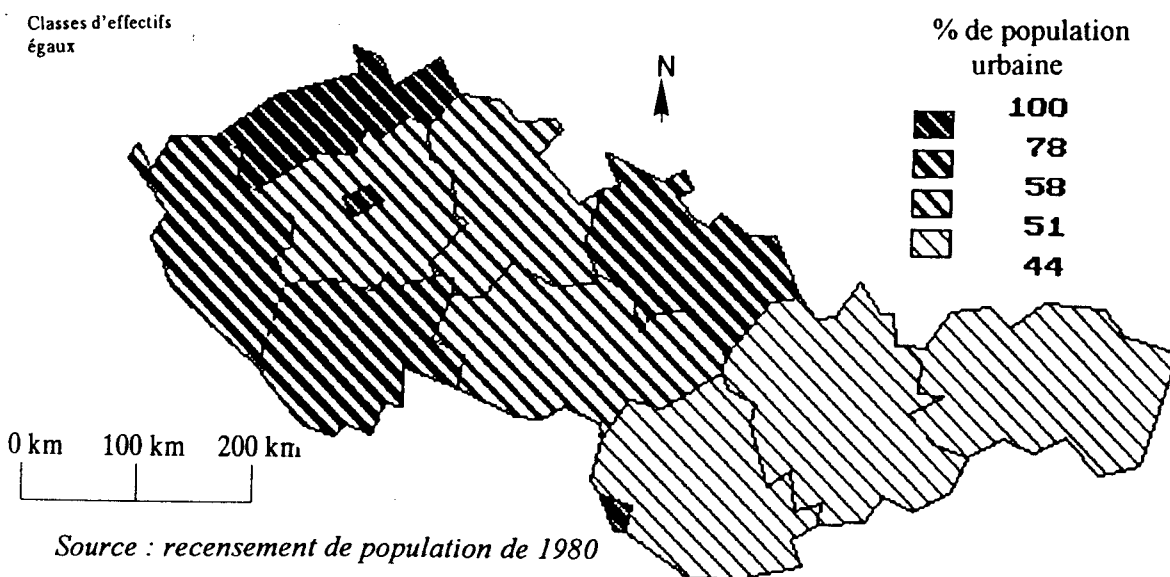
* CARTOGRAPHIE (20 pts)

2°) c) Les critères d'évaluation sont ceux utilisés d'habitude (choix des trames, qualité d'exécution, légende, sources, etc ...). Le sujet proposé comportait toutefois cinq difficultés spécifiques :

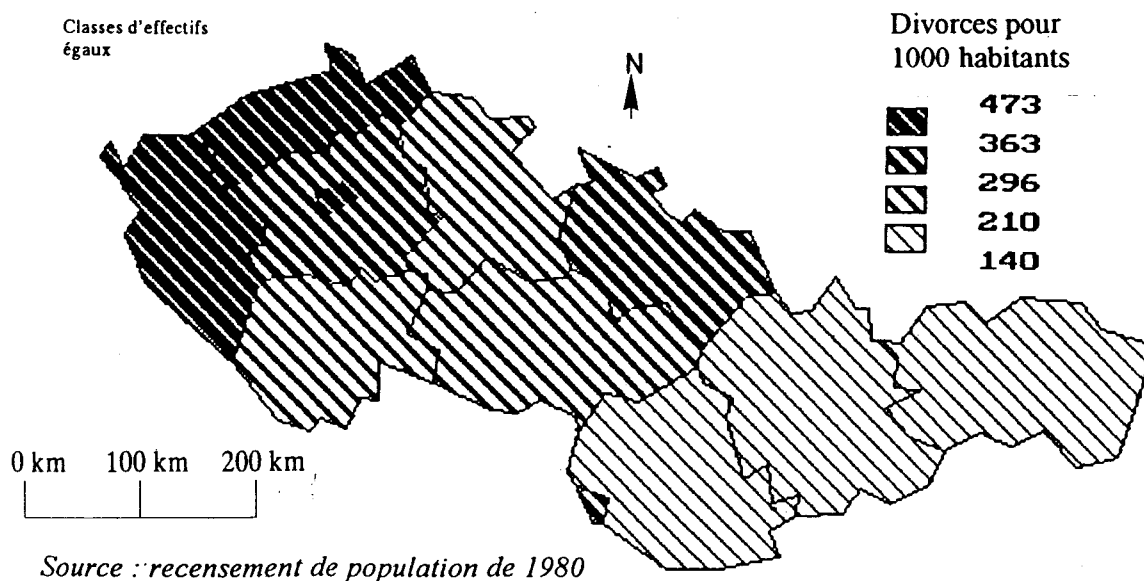
- une généralisation du tracé des contours de région devait être effectuée. Le fond proposé comporte en effet une précision exagérée dans le tracé des limites.
- la présence de deux unités spatiales de petite taille (Prague et Bratislava) imposait un choix raisonné des trames, afin que l'appartenance de ces unités à une classe ne soit pas ambiguë. Il fallait donc éviter des trames à motifs trop larges.
- l'objectif de comparabilité des deux cartes imposait nécessairement le choix des mêmes trames sur les deux cartes et une disposition identique des titres et des légendes.
- l'interdiction des à-plats noir et blanc visait à mesurer la capacité des étudiants à établir une progression visuelle régulière, du clair au foncé, avec quatre paliers.
- l'habillage de la carte devait être complet (titre, explication du mode de représentation cartographique, source, échelle) tout en utilisant le mieux possible les espaces blancs laissés par la forme du fond. En préservant un équilibre dans la disposition, il fallait essayer d'obtenir un format final minimal. En effet, il faut garder à l'esprit que l'on donne rarement beaucoup de place à la cartographie dans l'édition, même si on multiplie le nombre d'illustrations.

Remarque : l'exécution des deux cartes sur calque et à l'encre nécessite un apprentissage annuel. Le temps d'exécution proposé (environ 2 heures), n'est envisageable qu'avec des étudiants bien rodés à ce type d'exercice. On peut raccourcir l'exercice en fournissant les cartes des deux variables (page suivante) et en demandant aux étudiants de ne tracer que la carte des résidus. On peut aussi faire exécuter les cartes par les étudiants sur un logiciel de cartographie automatique.

TAUX D'URBANISATION DES REGIONS TCHECOSLOVAQUES EN 1980



TAUX DE DIVORCE DES REGIONS TCHECOSLOVAQUES EN 1980



*** RESULTATS FACE AUX ETUDIANTS**

129 étudiants inscrits en 1ere année de DEUG de Géographie à l'Université Paris 1 ont composé le sujet proposé.

	STATISTIQUES	CARTOGRAPHIE	
Minimum	1	0	
Maximum	18	16	
Mode	13	12	
Médiane	12	12	Le coefficient
Moyenne	11.4	11.3	de corrélation
Ecart-type	3.1	2.5	linéaire des deux
Notes ≤ 10	29 %	31 %	notations est de +0.41

Pour parodier le sujet de statistiques, on peut dire que ce tableau nous apprend que les deux distributions sont dissymétriques à droite, c'est-à-dire que la dispersion est plus forte dans les valeurs faibles que dans les valeurs fortes. Ce résultat s'explique par l'indépendance des questions (l'échec à une question ne compromettrait pas la réponse aux autres) et par la multiplicité des critères d'évaluation (à la fois du calcul et du commentaire). Dans l'ensemble, la dispersion des notes est plus forte en statistique qu'en cartographie. On observe toutefois une assez forte corrélation entre les deux séries de notes (+ 0.41). Cette corrélation s'explique par les liens étroits qui existent entre les deux enseignements. Il y a certes des déviations dans les valeurs moyennes (8 en statistiques et 12 en cartographie, ou l'inverse) mais les très bons ou les très mauvais résultats sont identiques en statistique et en cartographie.

Les réponses aux différentes questions, montrent une différence frappante entre la réussite générale des calculs statistiques, et la qualité très variable des commentaires géographiques déduits de ces calculs. Pour le commentaire de la carte, par exemple, il est rare que les étudiants décrivent à la fois la répartition géographique précise des deux phénomènes et les conséquences de la ressemblance entre ces deux répartitions (hypothèse d'une corrélation positive). Le plus souvent, un seul des aspects est présenté. Le problème est encore plus criant dans le cas du commentaire de l'analyse des résidus de régression. Trop souvent, les commentaires utilisent un "jargon" statistique mal maîtrisé et trop général (il pourrait s'appliquer à n'importe quelles variables). Cette difficulté de passer de résultats abstraits à un commentaire concret provient vraisemblablement d'un blocage de beaucoup d'étudiants d'origine littéraire vis-à-vis des statistiques. En effet, ce sont souvent les étudiants les plus hostiles aux aspects mathématiques de la statistique qui réalisent les commentaires les plus hermétiques.

Un autre type de problème a été révélé par les questions qui faisaient appel à une culture géographique générale et à une capacité de synthèse plutôt qu'à des outils précis. Les réponses sont apparues fréquemment incorrectes ou mal exprimées. Dans le cas de la question 1-b, très peu d'étudiants ont fait le lien entre le problème du choix de la variable et le cours sur les dangers inhérents à l'analyse des corrélations (effets de taille, effet de structure). Ceci montre la difficulté qu'ont les étudiants à faire des liens entre les différentes parties de l'enseignement, dès lors que ces liens n'ont pas été vus explicitement en Travaux Dirigés. Quant à la question 4-c, elle a rarement été l'occasion pour les étudiants de montrer les connaissances acquises dans d'autres U.V. (démographie, géographie humaine de l'Europe). Il y a manifestement là un problème de maturité, caractéristique d'étudiants qui sortent juste du secondaire.