Analyse de données sur le parcours des élèves en Bretagne à partir de données du rectorat

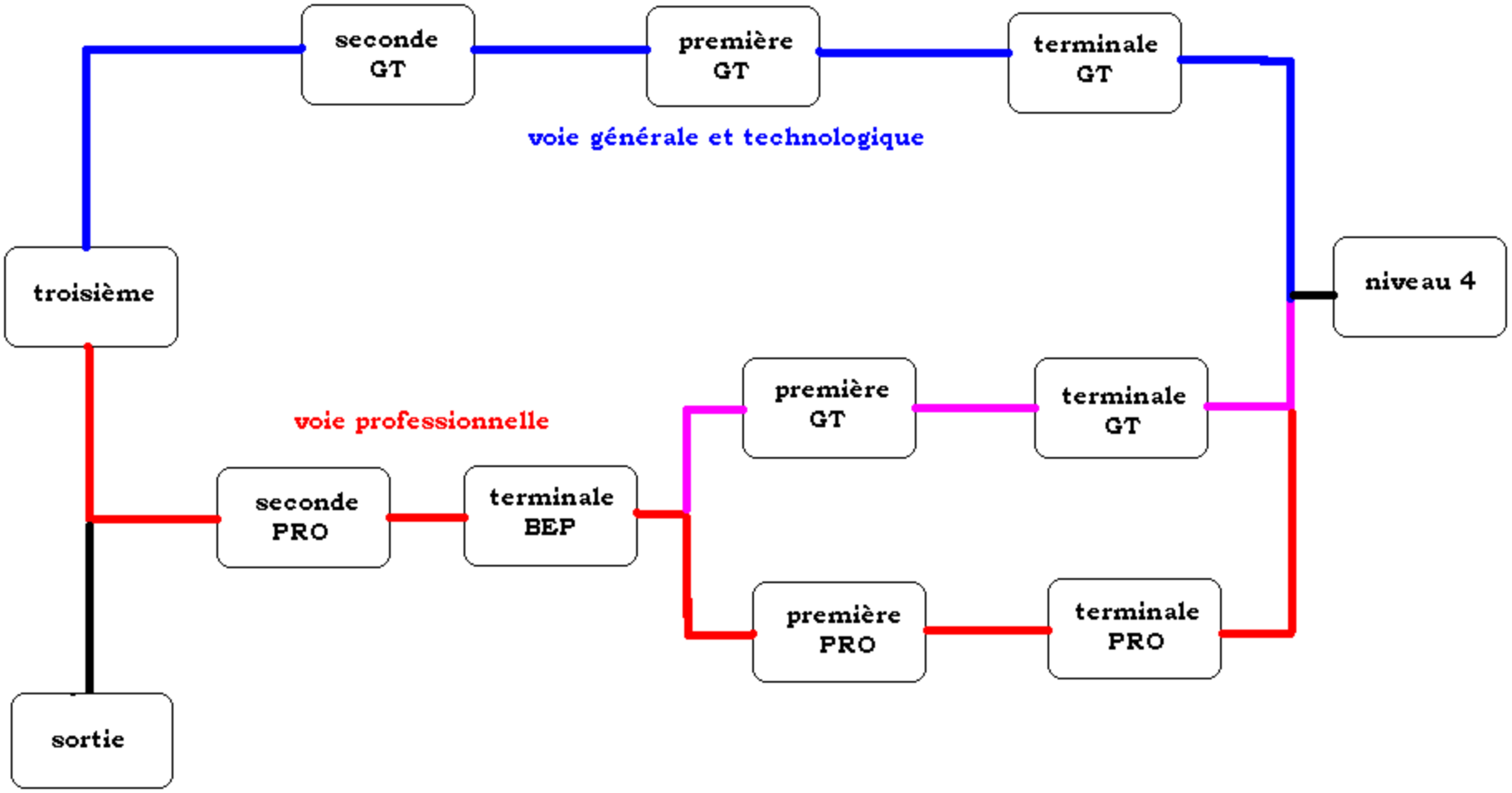
Stage fait à l'ENSAI

Du 14 mai au 6 juillet 2007

Les 12 Bassins d'Animation de la Politique Educative

- Auray-Ploërmel-Vannes
- Bain de Bretagne-Redon
- Brest
- Carhaix-Morlaix
- Combourg-Dinan-Saint Malo
- Fougères-Vitré
- Guingamp-Lannion
- Rennes
- Lorient-Quimperlé
- Pontivy-Loudéac
- Quimper
- Saint-Brieuc

Les différents parcours



Les 8 variables

- **Taux de 3ème net** : effectif de la cohorte de 3ème diminué des élèves sortis du système rapporté à l'ensemble de la cohorte;
- **Taux d'accès de la 3ème au niveau 4** : effectif atteignant le niveau 4 (quelque soit la voie et la durée) rapporté à l'effectif de 3ème net;
- **Part de la voie générale et technologique en seconde**;
- **Part de la voie GT au niveau 4**;
- **Taux de promotion de seconde GT au niveau 4** : effectif de terminale GT diminué de l'effectif passé par le BEP rapporté à l'effectif de 2nde GT;
- **Taux de fluidité du parcours GT** : effectif de terminale GT y accédant sans redoublement rapporté à l'effectif total accédant au niveau 4 GT hors effectif issu du BEP
- **Taux de promotion de 2nde professionnelle au niveau 4** : effectif de terminale professionnelle et de terminale GT issu du BEP rapporté à l'effectif de 2nde professionnelle
- **Taux de fluidité du parcours Pro** : effectif de terminale professionnelle et de terminale GT issu du BEP y accédant sans redoublement, rapporté à l'effectif total accéda,t au niveau par la voie professionnelle

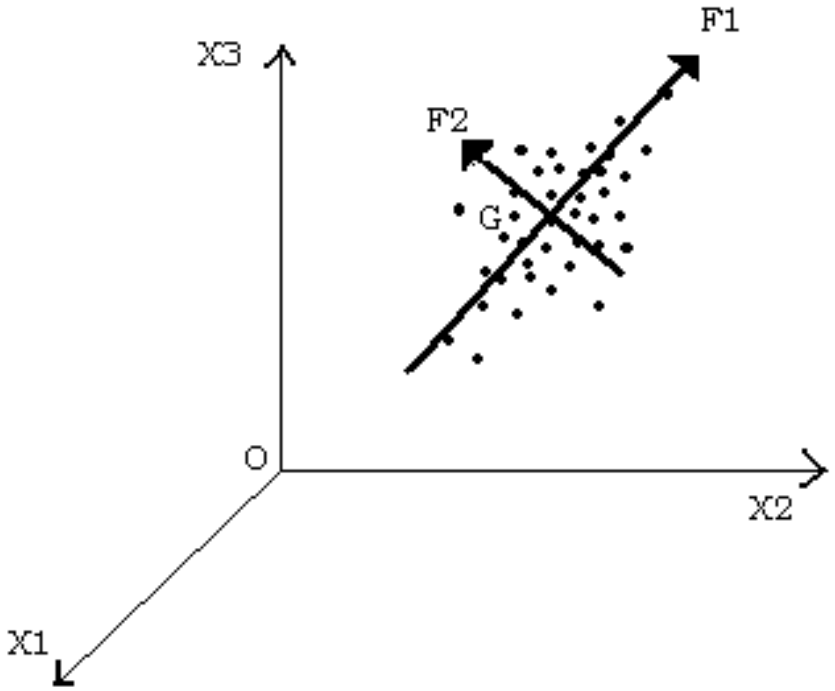
Le tableau des données

secteur	taux3net_99	acces3niv4_99	VoieGT_2_99	VoieGT_niv4_99	promGTniv4_99	fluidGT_99	promPRO_niv4_99	fluidPRO_99
AURAY/PLOERMEL/VANNES	84,71	77,74	66,49	84,13	87,51	63,60	57,62	69,54
BAIN DE BRETAGNE/REDON	82,34	77,84	67,57	85,65	87,73	68,69	58,52	74,76
BREST	89,67	79,08	74,18	92,25	87,28	65,93	52,15	64,39
CARHAIX/MORLAIX	83,56	77,10	67,03	89,18	88,38	74,77	53,54	71,37
COMBOURG/DINAN/ST-MALO	83,82	72,68	64,36	90,06	87,64	65,45	44,74	75,00
FOUGERES/VITRE	78,11	74,69	66,67	84,88	84,69	71,81	52,34	69,40
GUINGAMP/LANNION	89,02	80,38	69,96	91,90	90,80	61,64	55,14	74,09
LORIENT/QUIMPERLE	91,64	76,43	67,02	88,14	90,58	65,13	47,36	72,86
PONTIVY-LOUDEAC	83,14	72,14	58,67	92,92	91,29	67,92	47,43	80,83
QUIMPER	87,98	82,61	72,94	91,32	92,38	70,40	56,73	75,91
RENNES	88,55	82,29	75,66	92,87	91,86	71,91	52,58	69,63
SAINT-BRIEUC	88,72	80,68	70,67	92,22	92,13	68,62	51,72	72,96

Source : rectorat de Bretagne

Analyse en composantes principales

Principe :



- Déterminer un espace de faible dimension

Pour projeter sur ce sous-espace et réduire la vision du nuage dans un espace de dimensions plus faibles que l'espace de départ

Sous les contraintes :

conserver toute l'information utile

consentir une perte d'information minimale

- Pour déterminer le premier axe, on cherche à maximiser l'inertie du nuage projeté sur cet axe

U_1 : un vecteur directeur unitaire de l'axe Δu_1

X : la matrice de données de taille n×p

x_i : le vecteur de taille p×1 des données relatives à l'individu i

c_{i1} : la longueur de la projection de x_i sur l'axe Δu_1

$$c_{i1} = \langle x_i, u_1 \rangle$$

C_1 : le vecteur de taille n×1 de terme générique $C_1 = Xu_1$

On l'appelle le composante principale, formée des coordonnées des 12 secteurs sur l'axe 1

On a donc ainsi : $$\begin{aligned} I_1 &= {}^t C_1 C_1 \\ &= {}^t(Xu_1)(Xu_1) = {}^t u_1 {}^t X X u_1 \end{aligned}$$

Pour trouver u_1, on est donc conduit à résoudre le problème de minimisation suivant :

$$Max_{u_1}({}^t u_1 {}^t X X u_1) \text{ sous la contrainte } \|u_1\| = 1$$

- **Théorème 1** : Un vecteur unitaire qui caractérise le sous-espace à une dimension ajustant au mieux le nuage des n points individus dans R^p est ***un vecteur propre de la matrice d'inertie*** ***^tXX*** correspondant à la plus grande valeur propre.

- **Théorème 2** : Le sous-espace à q dimensions qui ajuste au mieux le nuage de n points individus dans R^p est engendré par ***q vecteurs propres orthogonaux de la matrice d'inertie*** ***^tXX*** correspondant aux q plus grandes valeurs propres.

Outils pour l'interprétation

On décide de centrer le nuage des individus

L'inertie : $I_j = \sum_{i=1}^{n} c_{ij}^2$

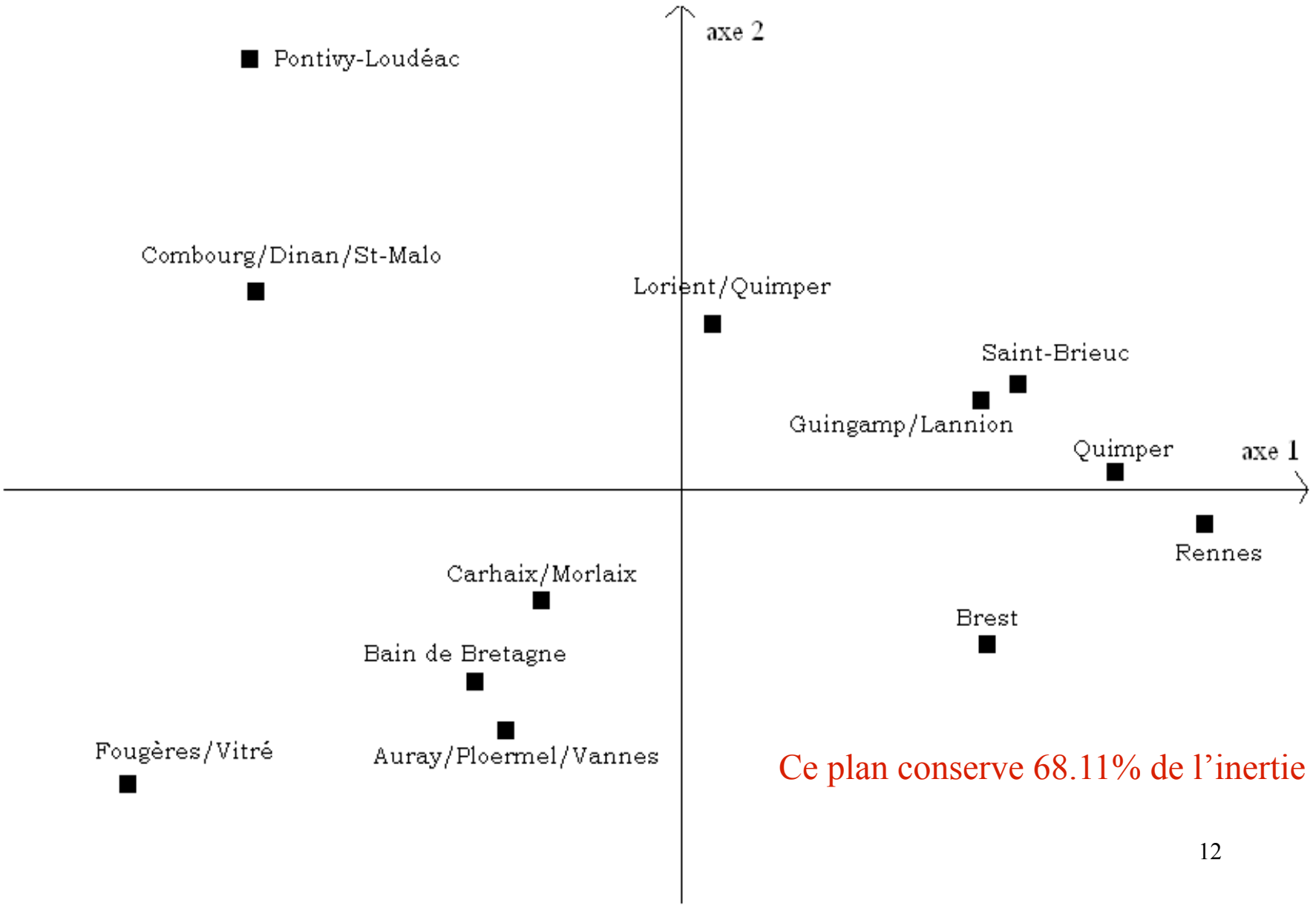
La contribution relative : $CTR_j(i) = \frac{c_{ij}^2}{I_j}$

La qualité de la représentation : $qlt_j(i) = cos^2\theta_{ij} = \frac{c_{ij}^2}{\|x_i\|^2}$

Aide à l'interprétation pour les individus actifs

Individus actifs			AXE 1				AXE 2			
Ident.	CONTR	POIDS	COORD	CTR	CO2	QLT	COORD	CTR	CO2	QLT
AURAY/PLOERMEL/VANNE	8.00	8.33	-0.88	2.0	10.0	10.0	**-1.77**	**11.6**	40.8	50.8
BAIN DE BRETAGNE/RED	5.84	8.33	-1.03	2.8	19.1	19.1	-1.41	7.4	35.7	54.8
BREST	9.26	8.33	1.54	6.2	26.6	26.6	-1.13	4.8	14.4	41.0
CARHAIX/MORLAIX	4.50	8.33	-0.70	1.3	11.5	11.5	-0.81	2.4	15.2	26.7
COMBOURG/DINAN/ST-MA	8.93	8.33	**-2.13**	**11.8**	52.9	52.9	1.48	8.1	25.6	78.6
FOUGERES/VITRE	14.00	8.33	**-2.78**	**20.1**	57.4	57.4	**-2.17**	**17.5**	35.2	92.5
GUINGAMP/LANNION	6.14	8.33	1.50	5.9	38.3	38.3	0.68	1.7	7.7	46.0
LORIENT/QUIMPERLE	5.50	8.33	0.16	0.1	0.5	0.5	1.24	5.7	29.1	29.6
PONTIVY-LOUDEAC	16.71	8.33	**-2.16**	**12.2**	29.2	29.2	**3.21**	**38.1**	64.2	93.4
QUIMPER	7.95	8.33	**2.18**	**12.3**	62.1	62.1	0.14	0.1	0.3	62.4
RENNES	9.29	8.33	**2.63**	**18.0**	77.3	77.3	-0.25	0.2	0.7	78.0
SAINT-BRIEUC	3.89	8.33	1.69	7.4	76.2	76.2	0.80	2.4	17.0	93.2

Les individus dans le plan Axe 1-Axe 2

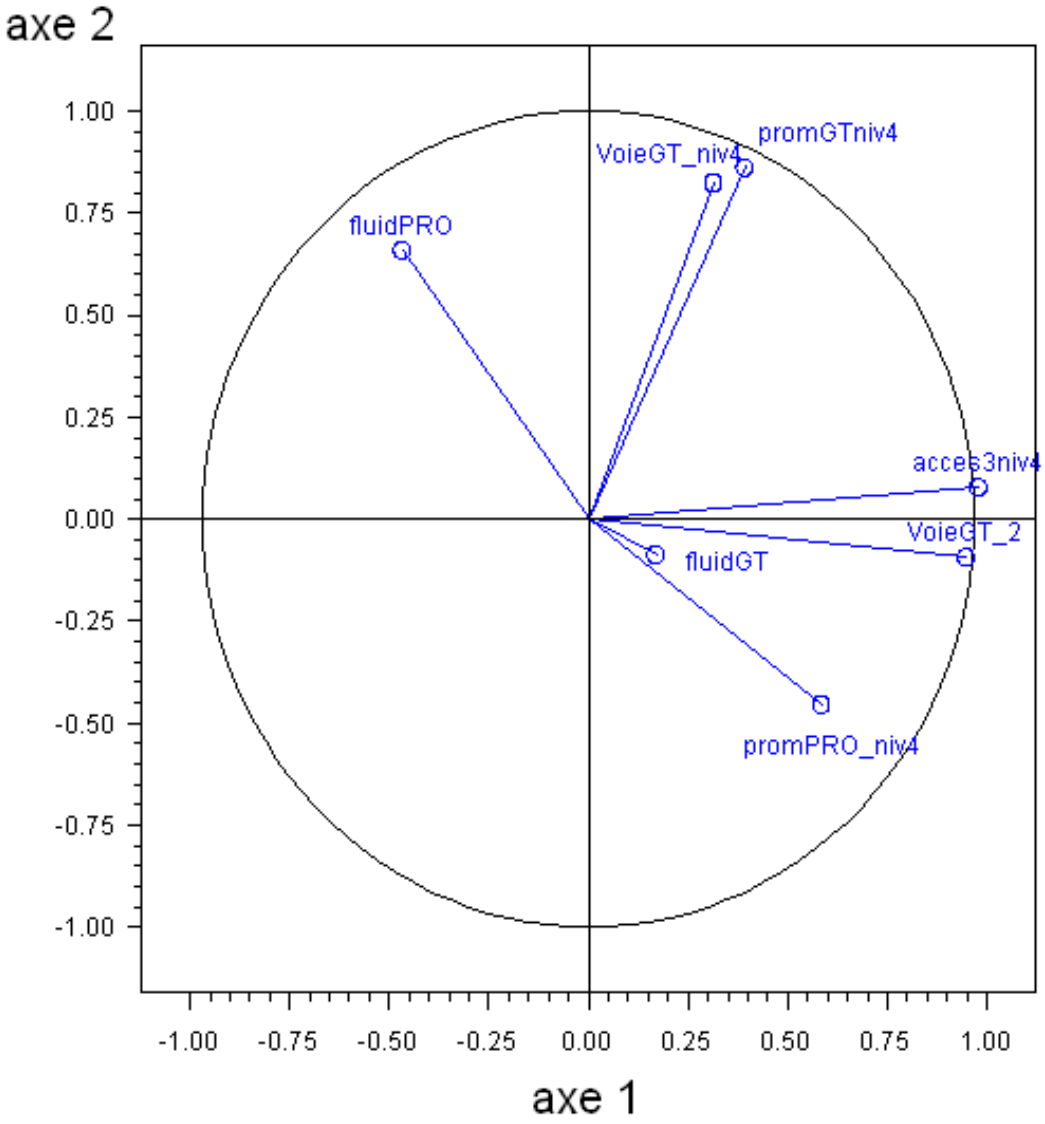


Ce plan conserve 68.11% de l'inertie

Aide à l'interprétation pour les variables actives

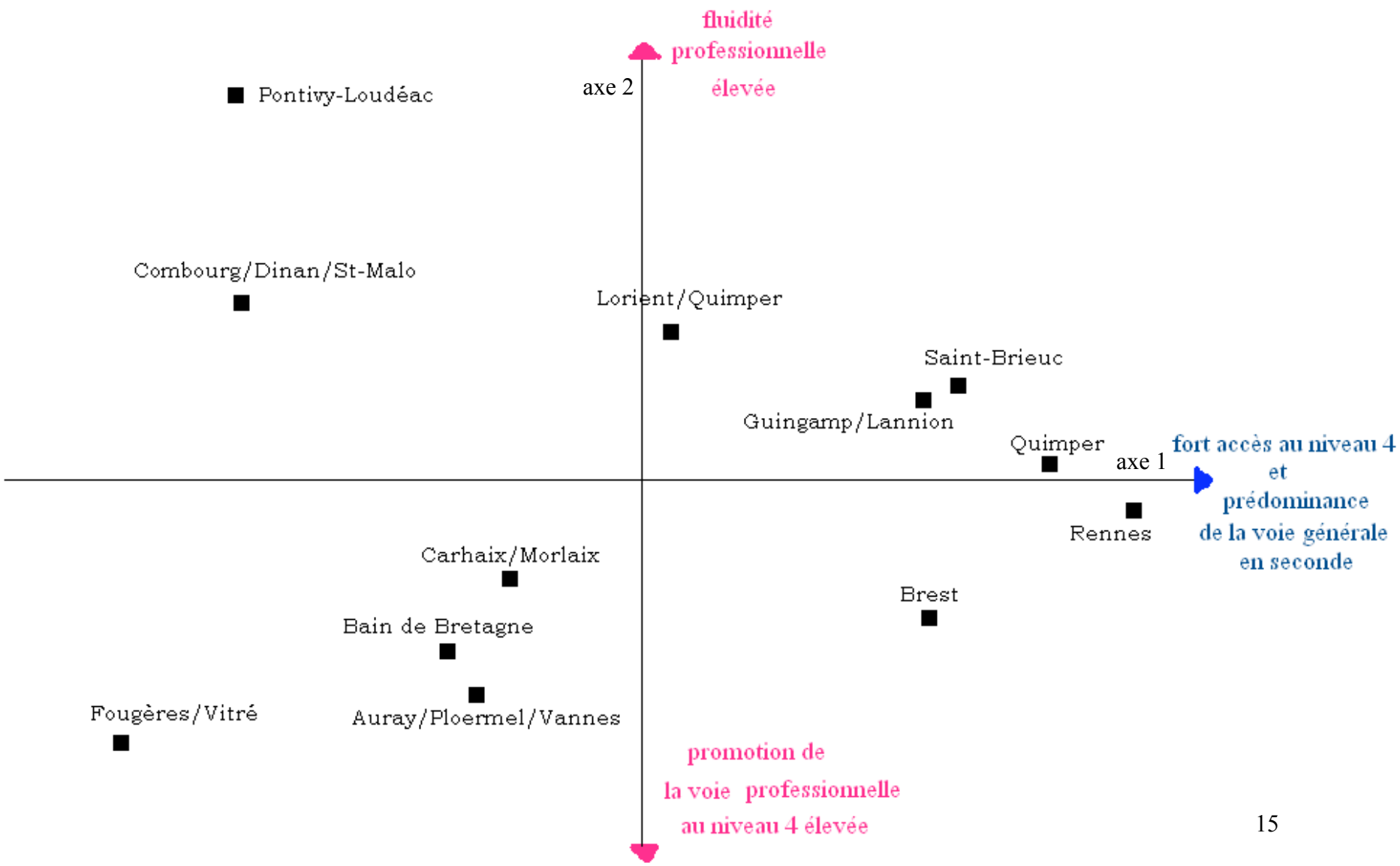
Variables actives			AXE 1				AXE 2			
Ident.	CONTR	POIDS	COORD	CTR	CO2	QLT	COORD	CTR	CO2	QLT
taux3net_99	12.50	12.50	**0.82**	**21.1**	67.5	67.5	0.31	4.3	9.6	77.1
acces3niv4_99	12.50	12.50	**0.92**	**26.5**	84.6	84.6	-0.30	4.1	9.1	93.8
VoieGT_2_99	12.50	12.50	**0.86**	**23.0**	73.6	73.6	-0.43	8.2	18.4	92.1
VoieGT_niv4_99	12.50	12.50	0.58	10.4	33.3	33.3	**0.64**	**18.1**	40.8	74.1
promGTniv4_99	12.50	12.50	0.66	13.7	43.8	43.8	**0.64**	**18.1**	40.7	84.5
fluidGT_99	12.50	12.50	-0.04	0.1	0.2	0.2	-0.25	2.8	6.4	6.5
promPRO_niv4_99	12.50	12.50	0.31	3.0	9.6	9.6	**-0.68**	**20.5**	46.1	55.7
fluidPRO_99	12.50	12.50	-0.27	2.3	7.3	7.3	**0.73**	**23.9**	53.9	61.2

Les variables dans le plan Axe 1-Axe 2

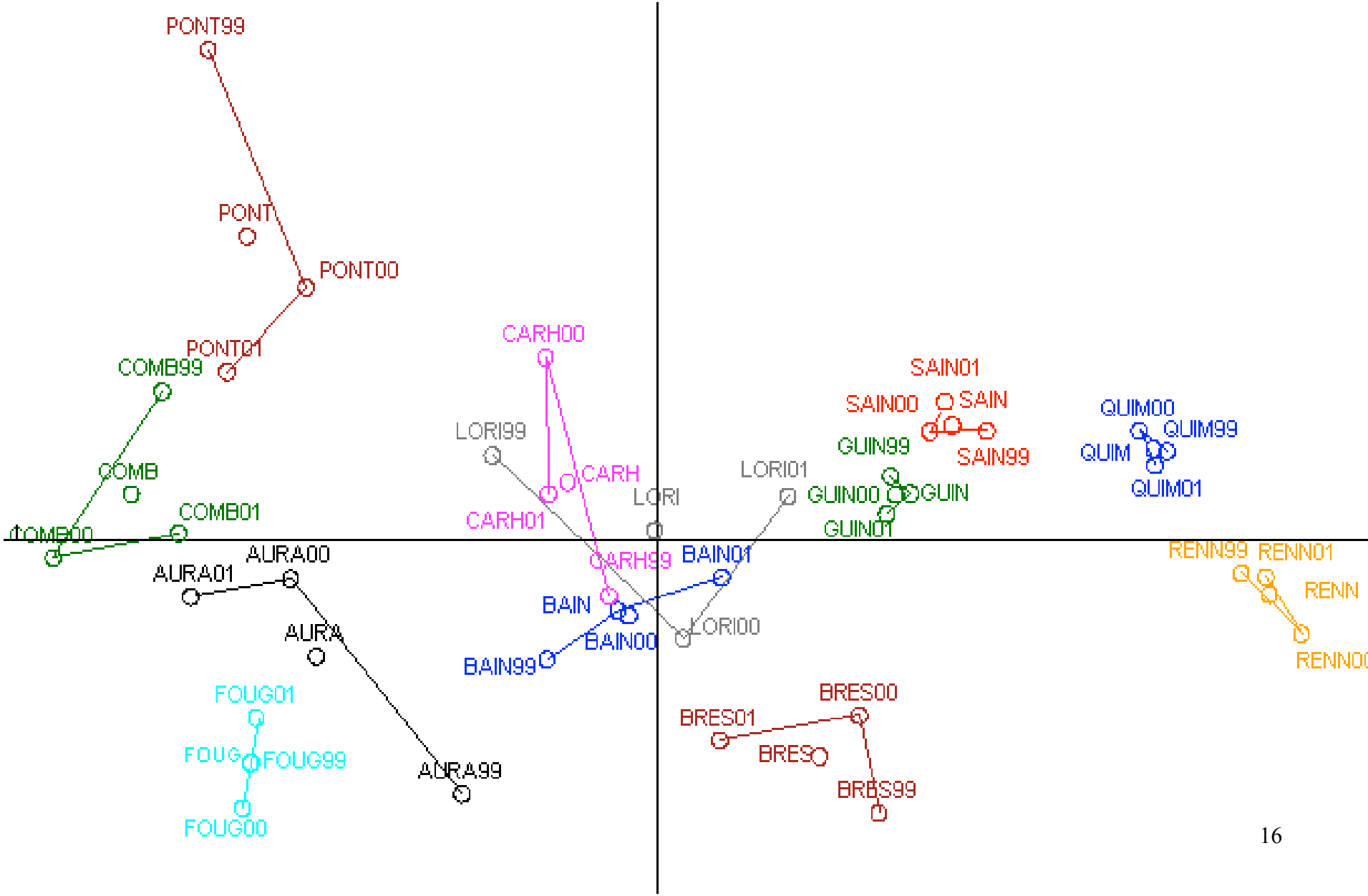


Corrélations entre variables

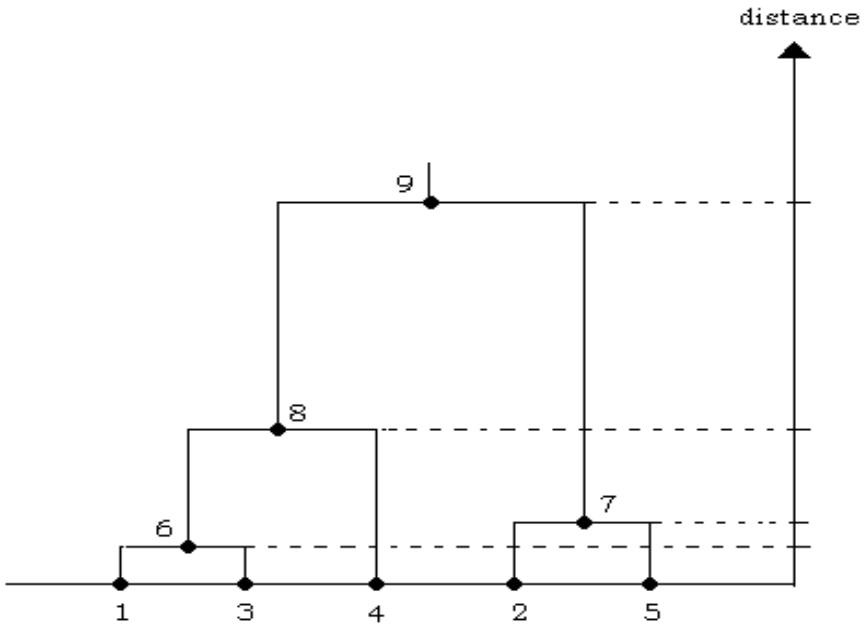
Interprétation



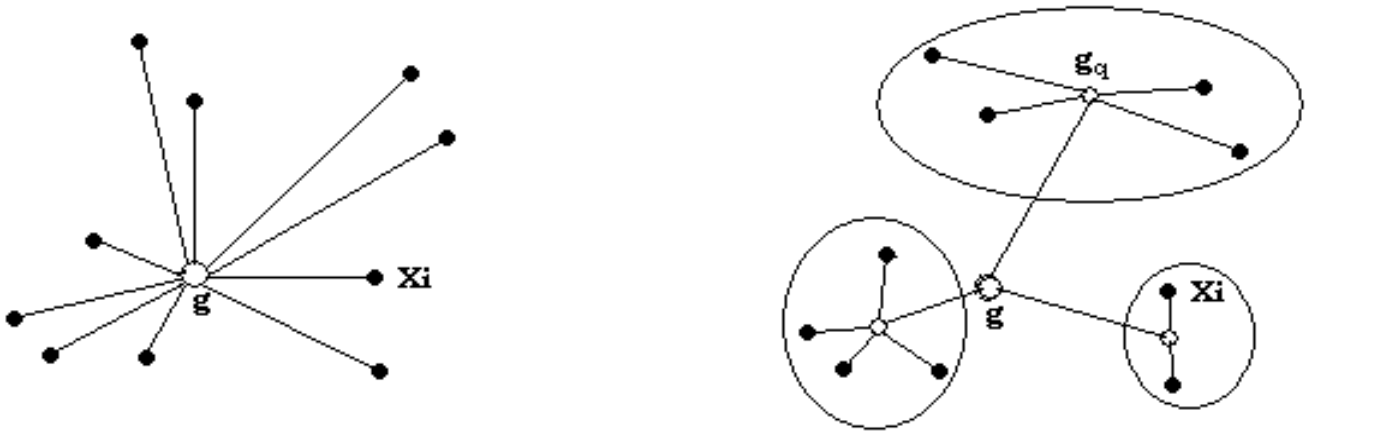
Analyse temporelle



Classification ascendante hiérarchique



La relation de Huygens :

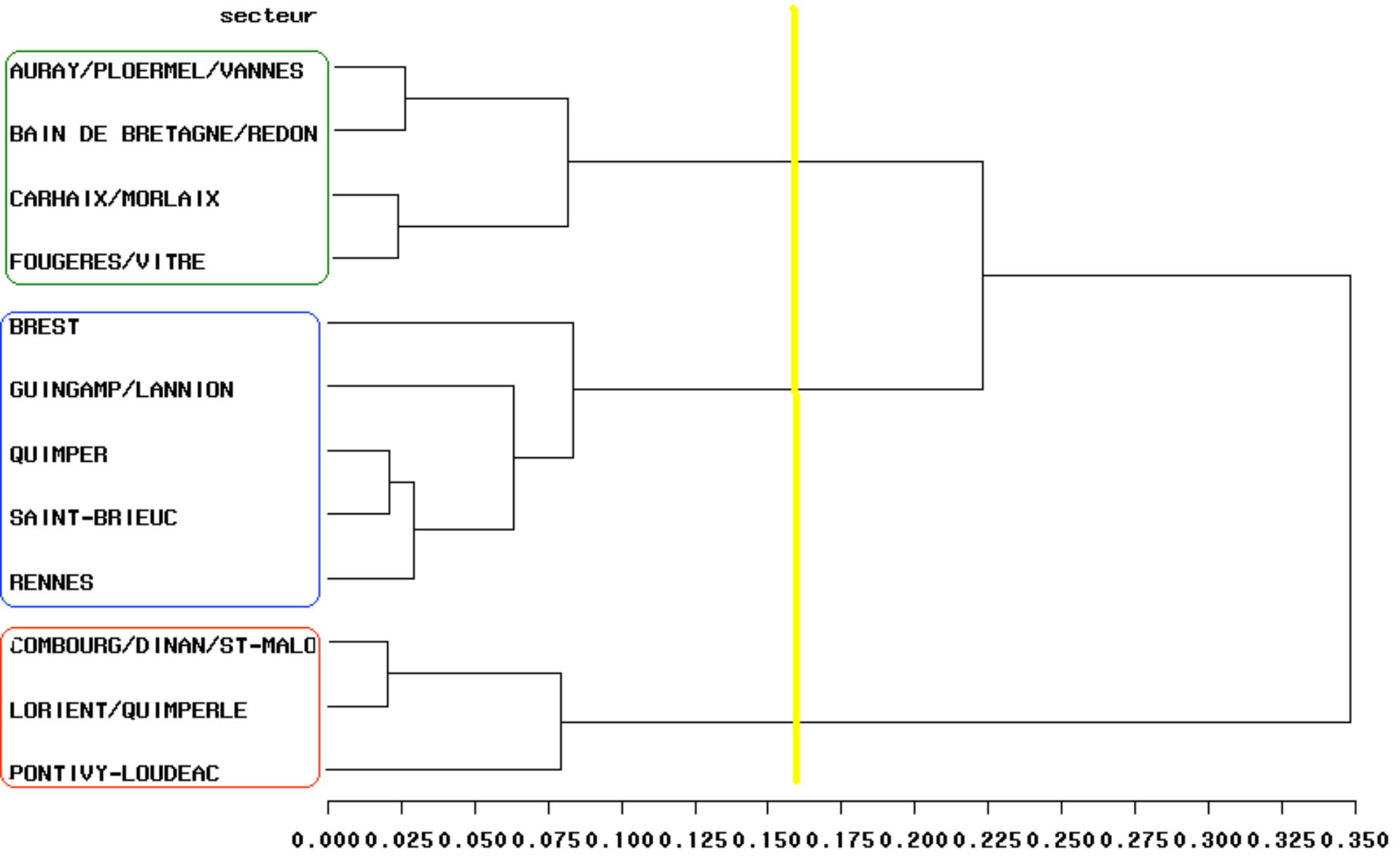


Inertie totale = Inertie inter-classes + Inertie intra-classes

Tableau des nœuds de la hiérarchie

Nœuds	Effectif	classes jointes	perte d'inertie inter			histogramme
	pondéré		0/00	cumul	diff	
CL1	12	CL2-CL3	317	317	0	**********
CL2	7	CL4-CL5	251	568	67	**********
CL3	5	BREST-CL6	94	662	157	**********
CL4	4	CL9-CL7	78	740	15	**********
CL5	3	CL8-PONTI	72	812	6	**********
CL6	4	GUING-CL10	61	874	11	**********
CL7	2	CARHA-FOUGE	35	909	26	**********
CL8	2	COMBO-LORIE	26	934	10	**********
CL9	2	AURAY-BAIN	24	958	2	**********
CL10	3	CL11-RENNE	24	982	0	**********
CL11	2	QUIMP-SAINT	18	1000	6	**********

Arbre de classification



On obtient donc trois classes :

Classe 1 : Rennes, Brest, Quimper, St Brieuc, Guingamp-Lannion

Classe 2 : Auray-Ploërmel-Vannes, Bain de Bretagne-Redon, Carhaix-Morlaix, Fougères-Vitré

Classe 3 : Combourg-Dinan-Saint Malo, Lorient-Quimperlé, Pontivy-Loudéac

hypothèse H_0 : $\overline{X_k} = \overline{X}$

La statistique du test vaut :

$$t_k(X) = \frac{\overline{X_k} - \overline{X}}{\sigma_k(X)}$$

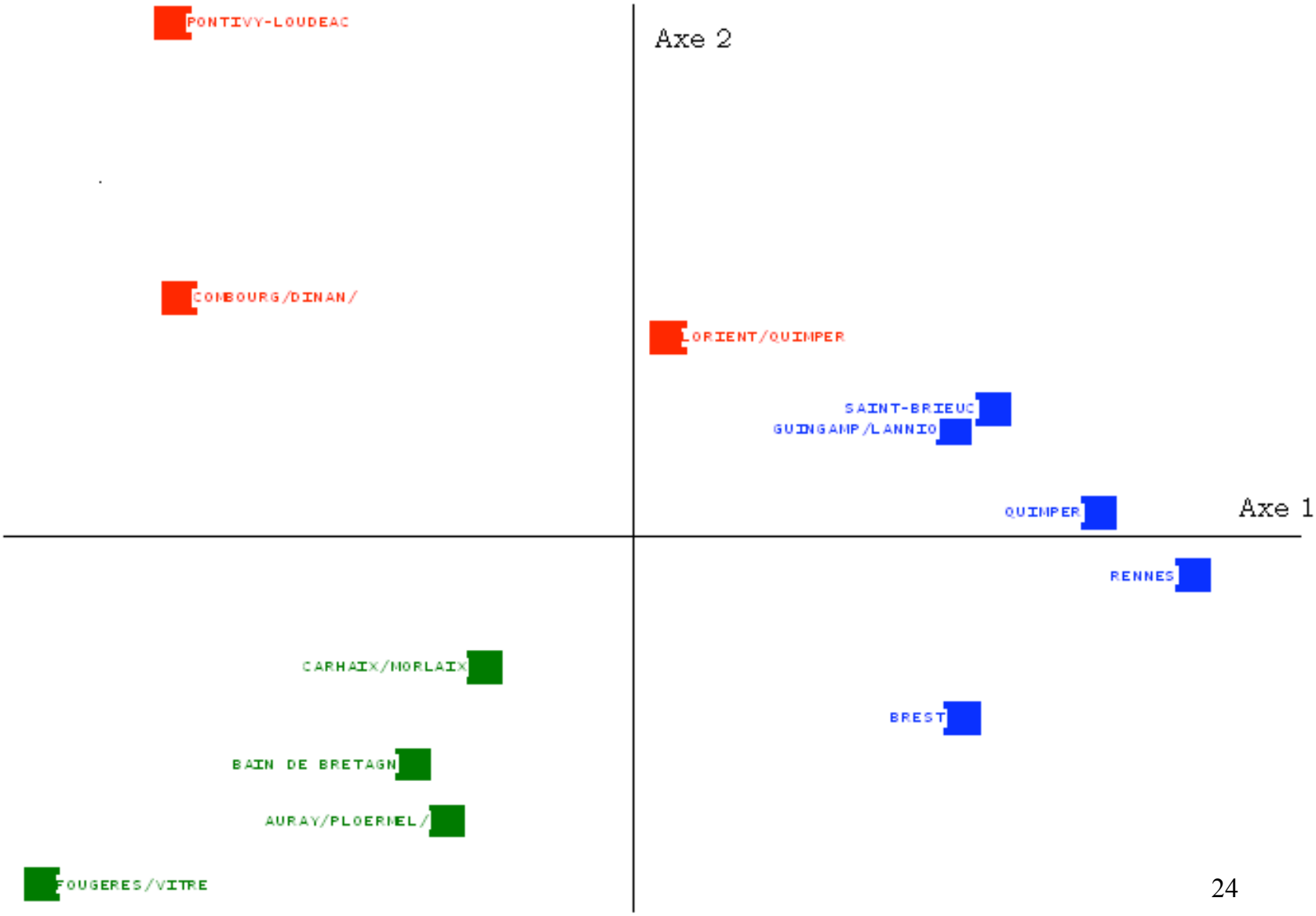
où $\overline{X_k}$ est la moyenne d'une variable X dans la classe k, $\overline{X}$ la moyenne générale de X et $\sigma_k(X)$ la variance de cette variable dans la classe.

Sous H_0, la statistique du test suit une loi normale centrée réduite. On rejette H_0 si $|t_k|$ est élevé, par exemple $|t_k| > 1.96$ (pour un niveau de significativité du test de 5%)

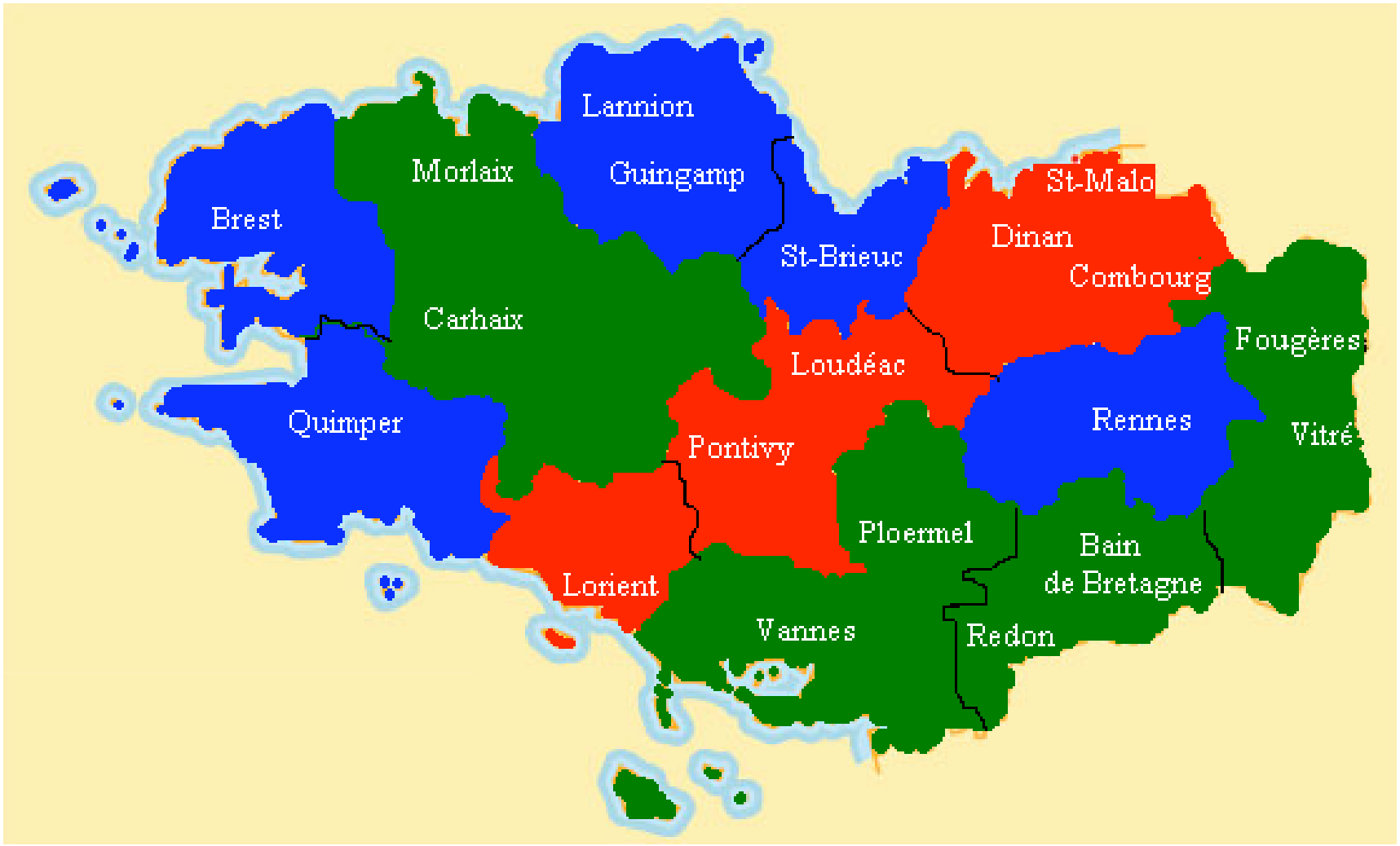
Description des classes

Variable	Statistique	_1	_2	_3
	Eff.pond	5.0000	4.0000	3.0000
VoieGT_2_99	Moyenne	72.6841	66.9386	63.3486
	Ec.type	2.1294	0.4133	3.4825
	V.test	**2.6824**	-0.7904	**-2.1936**
	Proba	0.0073	0.4293	0.0283
VoieGT_niv4_99	Moyenne	92.1126	85.9616	90.3726
	Ec.type	0.5049	1.9358	1.9631
	V.test	**2.2590**	**-2.7876**	0.4628
	Proba	0.0239	0.0053	0.6435
acces3niv4_99	Moyenne	81.0101	76.8435	73.7489
	Ec.type	1.2963	1.2725	1.9106
	V.test	**2.7262**	-0.6851	**-2.3580**
	Proba	0.0064	0.4933	0.0184
fluidGT_99	Moyenne	67.7019	69.7166	66.1637
	Ec.type	3.6268	4.1340	1.2482
	V.test	-0.2195	1.1057	-0.9538
	Proba	0.8262	0.2689	0.3402
fluidPRO_99	Moyenne	71.3969	71.2661	76.2319
	Ec.type	4.0577	2.1595	3.3687
	V.test	-0.8231	-0.7660	**1.7710**
	Proba	0.4105	0.4437	0.0766
promGTniv4_99	Moyenne	90.8904	87.0797	89.8368
	Ec.type	1.8840	1.4140	1.5796
	V.test	1.8272	**-2.2698**	0.3907
	Proba	0.0677	0.0232	0.6960
promPRO_niv4_99	Moyenne	53.6634	55.5071	46.5104
	Ec.type	1.9398	2.6202	1.2513
	V.test	0.8028	**1.7267**	**-2.7938**
	Proba	0.4221	0.0842	0.0052

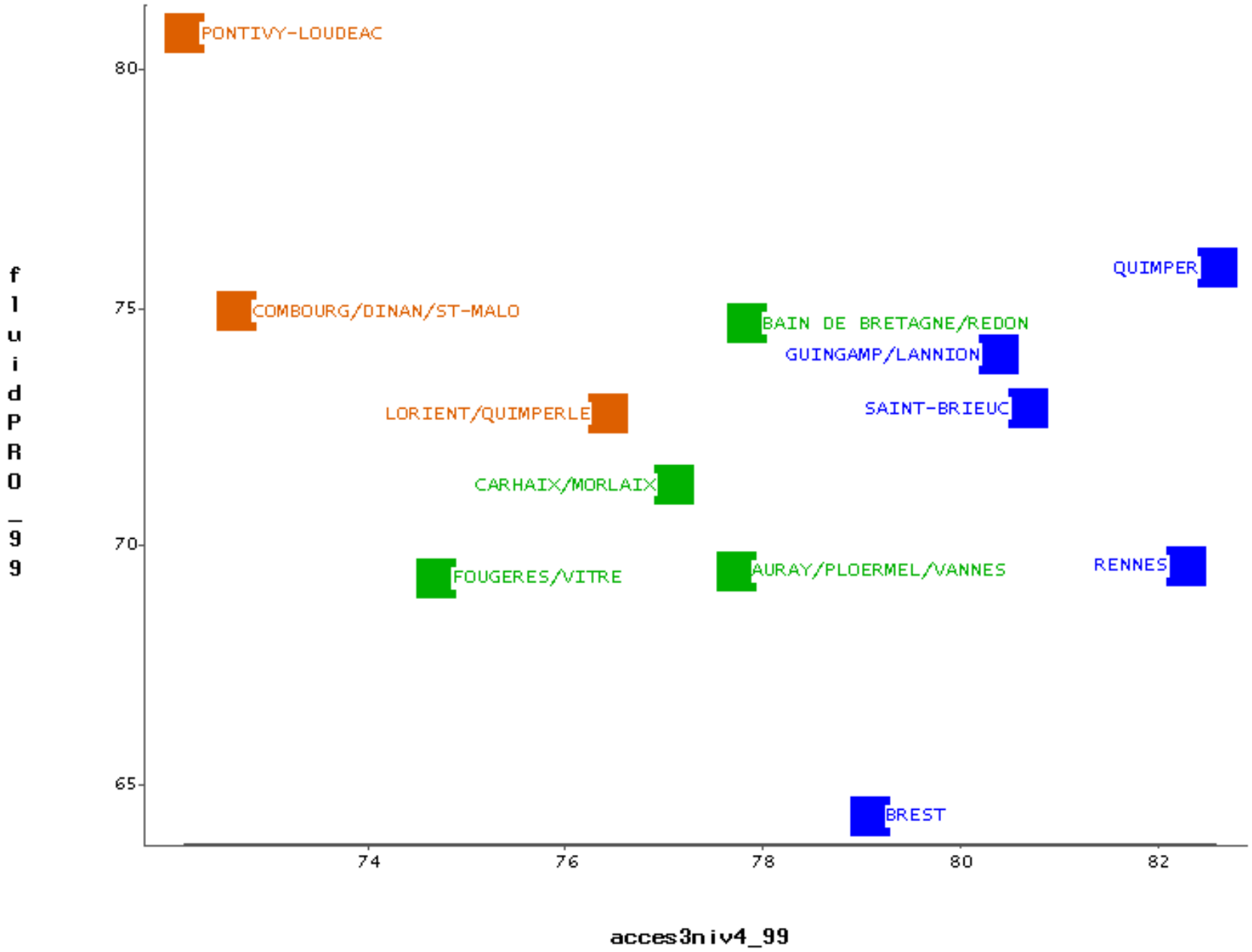
Visualisation des classes sur l'ACP

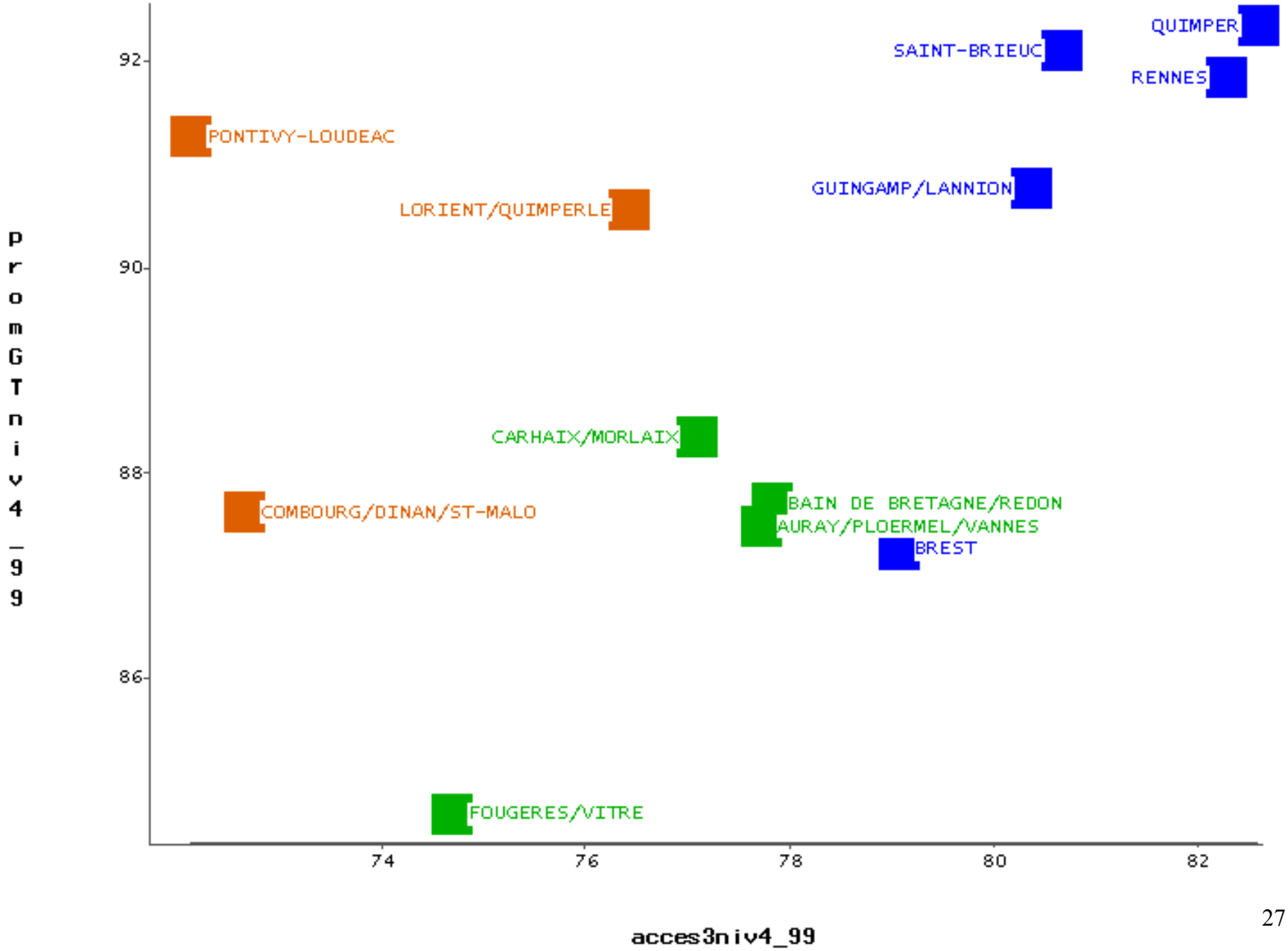


Carte géographique des classes



Visualisation des classes suivant différentes variables




QUIMPER
SAINT-BRIEUC
RENNES
PONTIVY-LOUDEAC
GUINGAMP/LANNION
LORIENT/QUIMPERLE
CARHAIX/MORLAIX
COMBOURG/DINAN/ST-MALO
BAIN DE BRETAGNE/REDON
AURAY/PLOERMEL/VANNES
BREST
FOUGERES/VITRE
92
90
88
86
74
76
78
80
82
promGTniv4_99
acces3niv4_99