ANR CORPUS GEOMEDIA

Paris, 24-25 novembre 2014

Agrégation multiéchelle de l'information médiatique

Robin Lamarche-Perrin
Jean-Marc Vincent
Yves Demazeau

L I G

Max Planck Institute for Mathematics in the Sciences

Analyse médiatique des relations internationales

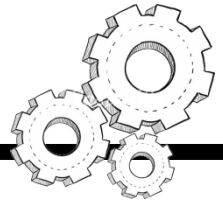
Hypothèse : les médias constituent un instrument d'observation adéquat au niveau national

Expert

Système international

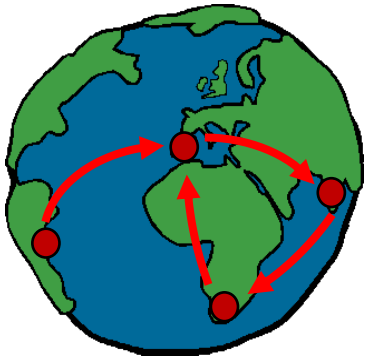
Media

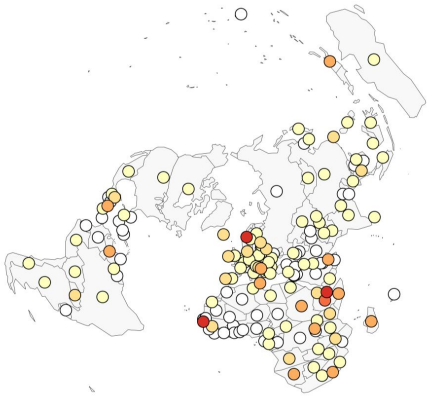
Représentation microscopique



Agrégation de données

Représentation macroscopique





Représentation microscopique du système international

Journal LE MONDE

π **Espace**

	USA	Libye	Syrie	France	Israël	…	Total
2 mai	25	12	11	10	4	…	**142**
9 mai	14	6	12	12	5	…	**108**
16 mai	20	11	12	6	9	…	**142**
23 mai	15	9	6	13	5	…	**120**
30 mai	10	16	17	9	4	…	**137**
6 juin	14	16	16	9	4	…	**114**
13 juin	15	14	17	9	6	…	**119**
20 juin	17	13	12	12	7	…	**123**
27 juin	7	6	7	20	2	…	**103**
4 juill.	12	13	8	10	6	…	**129**
11 juill.	21	10	10	14	3	…	**107**
18 juill.	7	3	8	4	5	…	**61**
25 juill.	16	7	6	13	4	…	**128**
1 août	21	1	9	7	4	…	**88**
…	…	…	…	…	…	…	**…**
Total	**423**	**308**	**260**	**248**	**153**	**…**	**3520**

t **Temps**

Nombre de citations observées

$v(\pi, t)$

$v(., t)$

$v(\pi, .)$

$v(., .)$

Nombre de citations attendues

$$v^*(\pi, t) = \frac{v(\pi, .)\, v(., t)}{v(., .)}$$

Représentation géographique

Espace

Temps

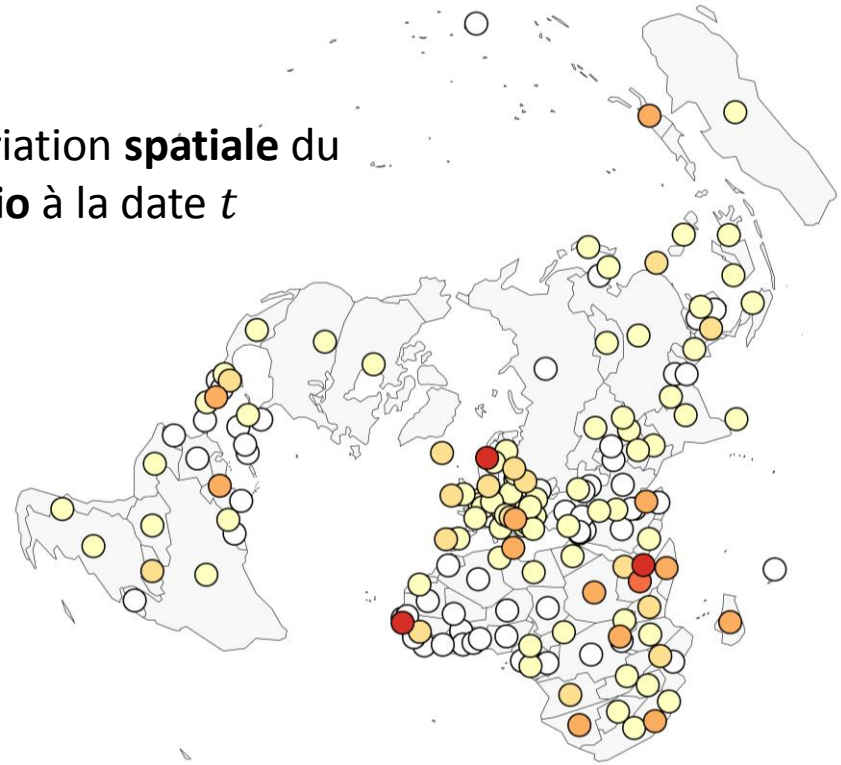
t

	USA	Libye	Syrie	France	Israël	…	Total
2 mai	25	12	11	10	4	…	142
9 mai	14	6	12	12	5	…	108
16 mai	20	11	12	6	9	…	142
23 mai	15	9	6	13	5	…	120
30 mai	10	16	17	9	4	…	137
6 juin	14	16	16	9	4	…	114
13 juin	15	14	17	9	6	…	119
20 juin	17	13	12	12	7	…	123
27 juin	7	6	7	20	2	…	103
4 juill.	12	13	8	10	6	…	129
11 juill.	21	10	10	14	3	…	107
18 juill.	7	3	8	4	5	…	61
25 juill.	16	7	6	13	4	…	128
1 août	21	1	9	7	4	…	88
…	…	…	…	…	…	…	…
Total	423	308	260	248	153	…	3520

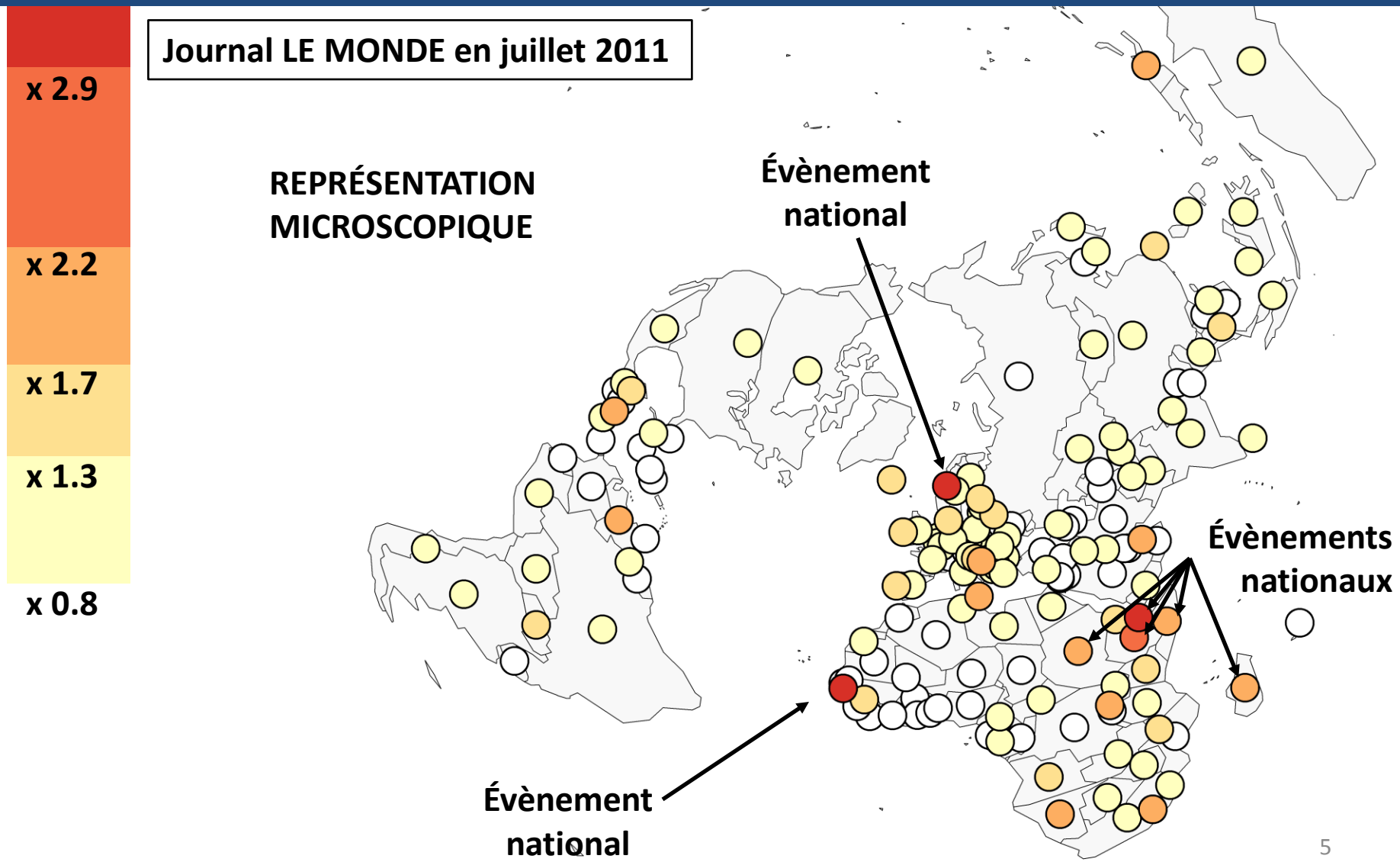
Ratio des citations observées et des citations attendues

$$\rho(\pi,t) = \frac{v(\pi,t)}{v^*(\pi,t)} = \frac{v(\pi,t)\; v(.,.)}{v(\pi,.)\; v(.,t)}$$

Variation **spatiale** du **ratio** à la date t



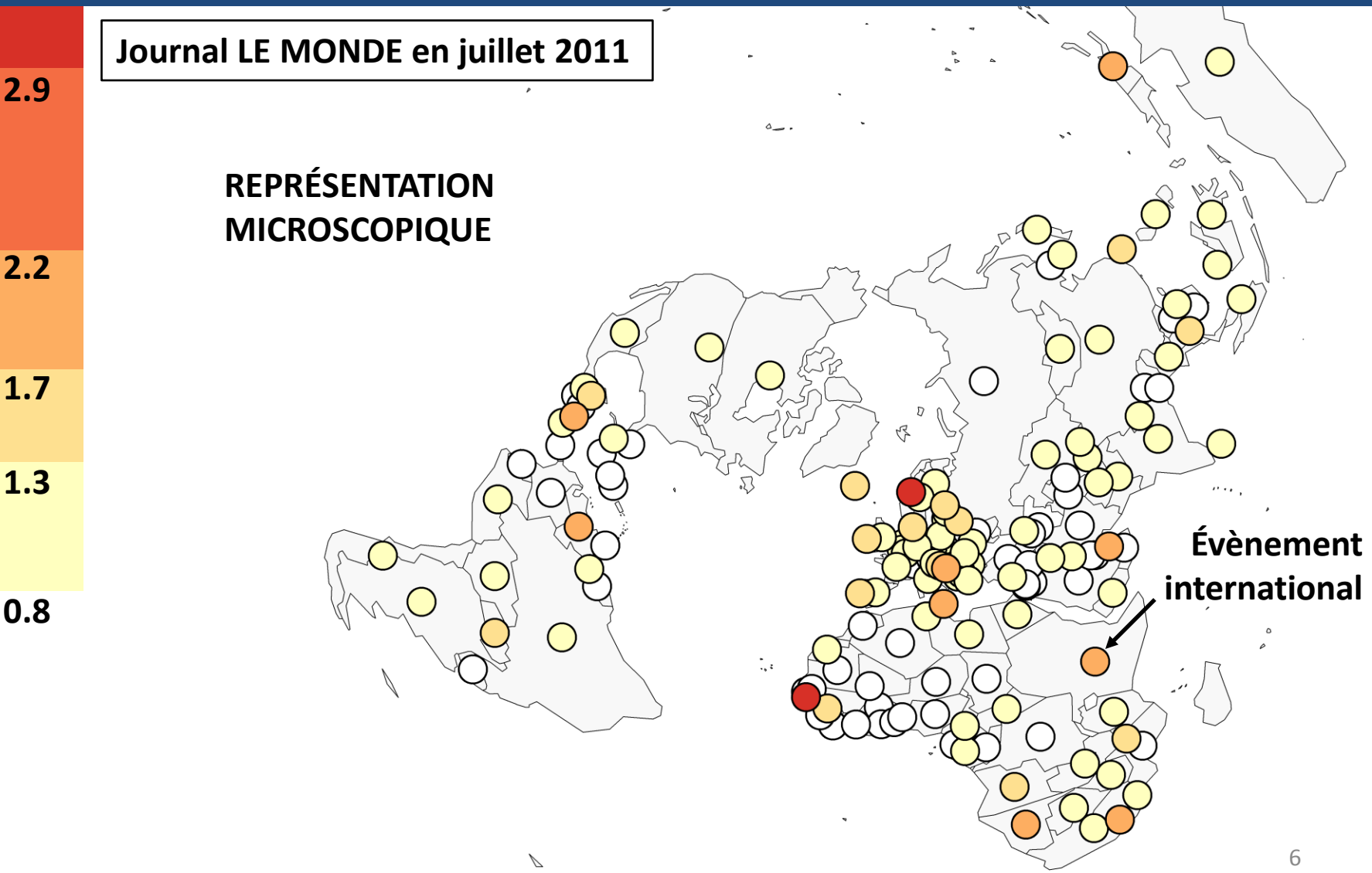
Détection d'évènements médiatiques



Détection d'évènements médiatiques

Journal LE MONDE en juillet 2011

REPRÉSENTATION MICROSCOPIQUE



Agrégation de données

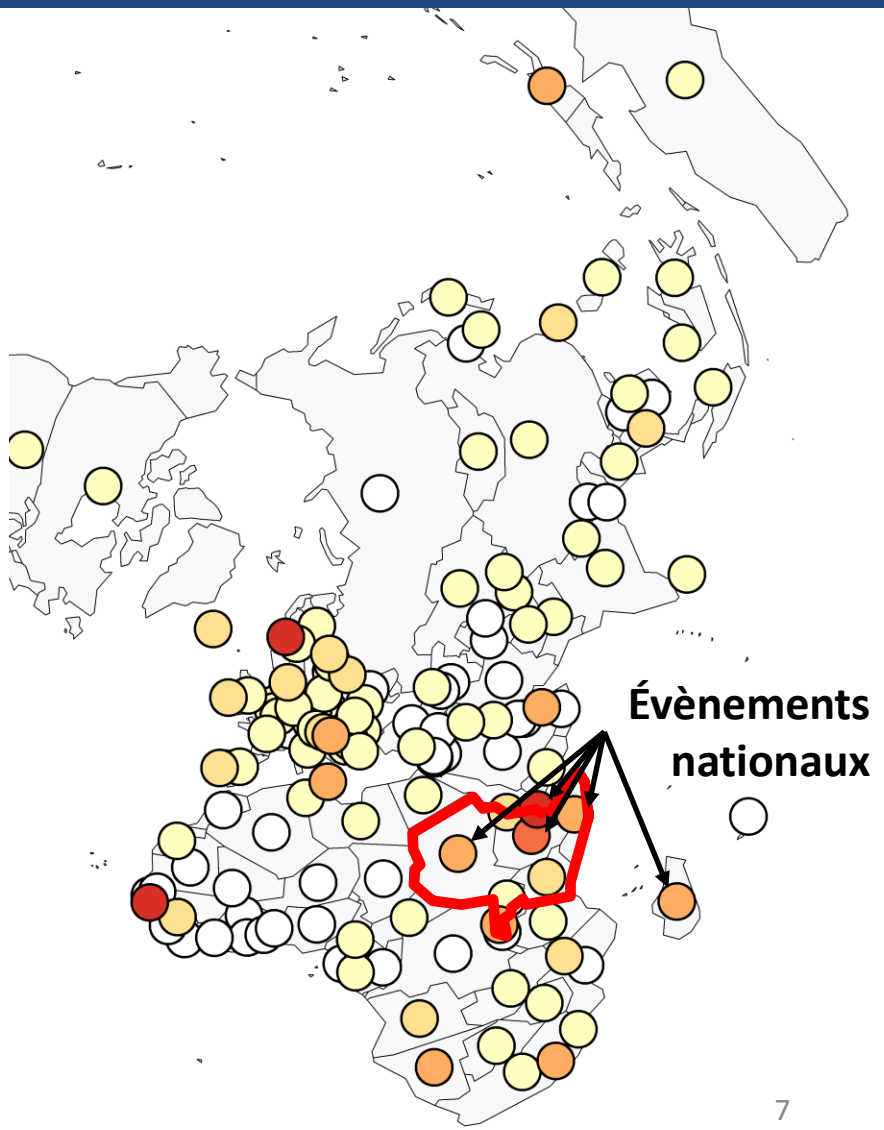
Espace

π_1 π_2 π_3

Temps

t

	USA	Libye	Syrie	France	Israël	...	Total
2 mai	25	12	11	10	4	...	142
9 mai	14	6	12	12	5	...	108
16 mai	20	11	12	6	9	...	142
23 mai	15	9	6	13	5	...	120
30 mai	10	16	17	9	4	...	137
6 juin	14	16	16	9	4	...	114
13 juin	15	14	17	9	6	...	119
20 juin	17	13	12	12	7	...	123
27 juin	7	6	7	20	2	...	103
4 juill.	12	13	8	10	6	...	129
11 juill.	21	10	10	14	3	...	107
18 juill.	7	3	8	4	5	...	61
25 juill.	16	7	6	13	4	...	128
1 août	21	1	9	7	4	...	88
...	...	...	...	...	...	...	...
Total	423	308	260	248	153	...	3520

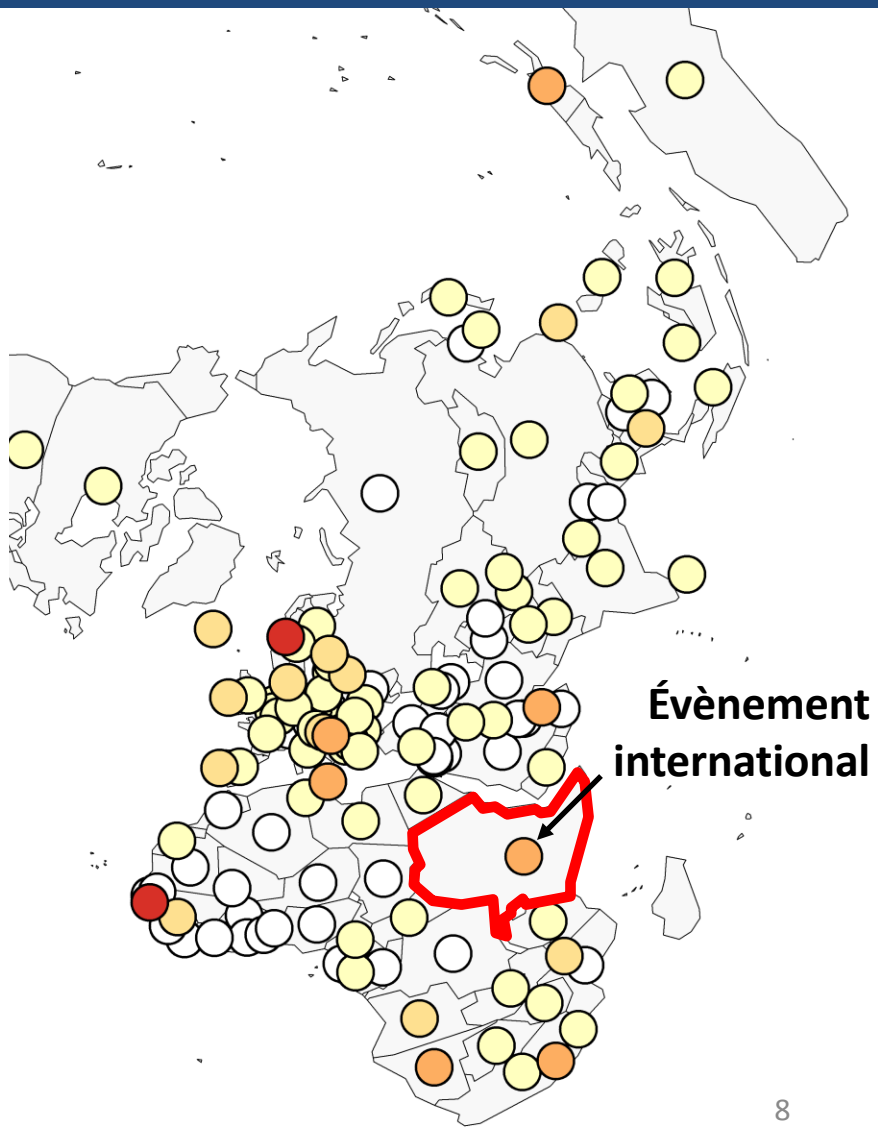


Agrégation de données

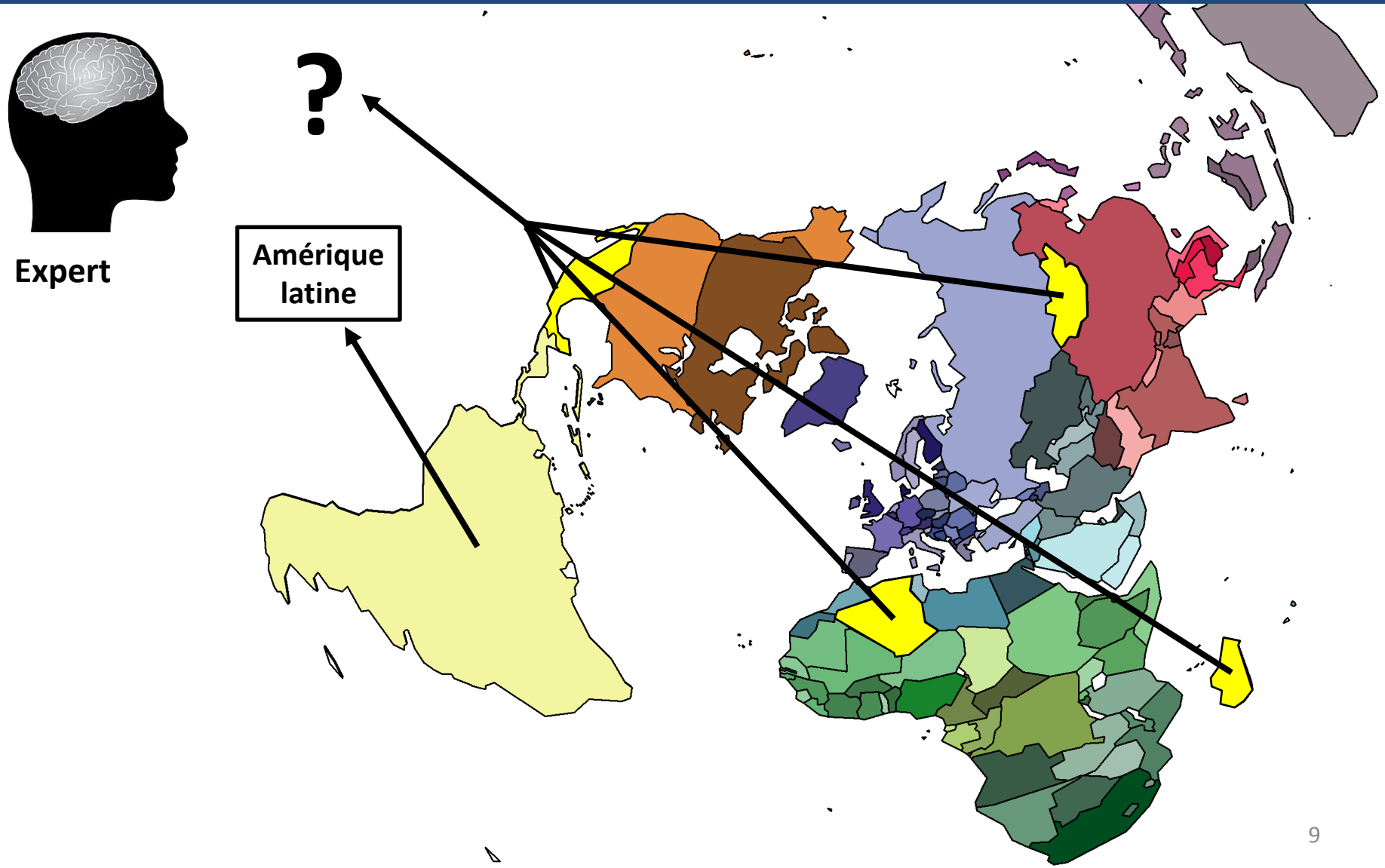
Espace π_1 π_2 π_3

Temps t

	USA	Agrégat	Israël	...	Total
2 mai	25	13+11+10	4	...	142
9 mai	14	6+12+12	5	...	108
16 mai	20	11+12+6	9	...	142
23 mai	15	9+6+13	5	...	120
30 mai	10	16+17+9	4	...	137
6 juin	14	16+16+9	4	...	114
13 juin	15	14+17+9	6	...	119
20 juin	17	13+12+12	7	...	123
27 juin	7	6+7+20	2	...	103
4 juill.	12	13+8+10	6	...	129
11 juill.	21	10+10+14	3	...	107
18 juill.	7	3+8+4	5	...	61
25 juill.	16	7+6+13	4	...	128
1 août	21	1+9+7	4	...	88
...	...	...	...	...	...
Total	423	308+260+248	153	...	3520



P1 : Sémantique géographique des agrégats



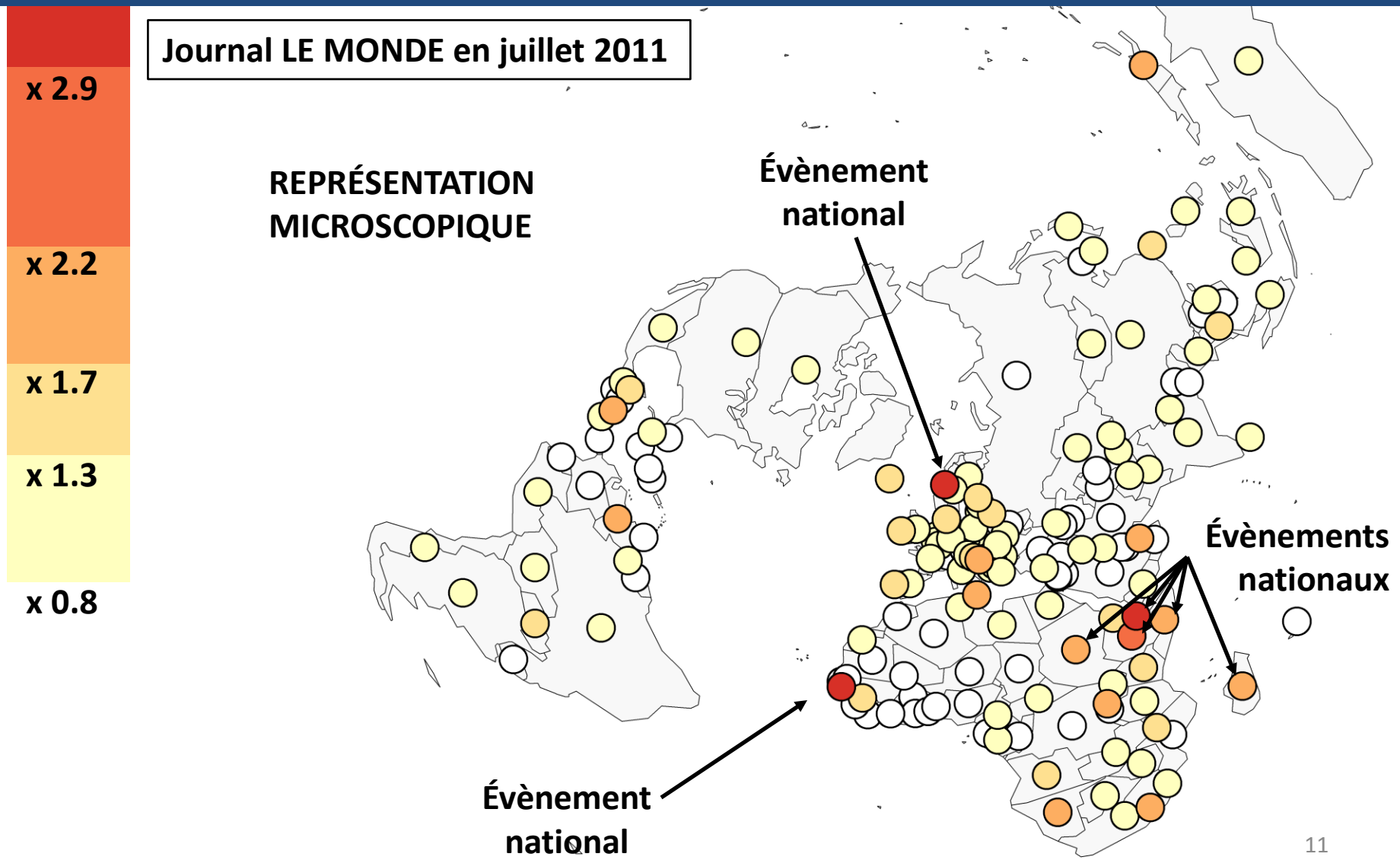
P1 : Sémantique géographique des agrégats

Expert

?

Problème 1 : Comment engendrer des abstractions cohérentes avec l'espace géographique ?

P2 : Niveaux de représentation



P2 : Niveaux de représentation

Journal LE MONDE en juillet 2011

x 2.9

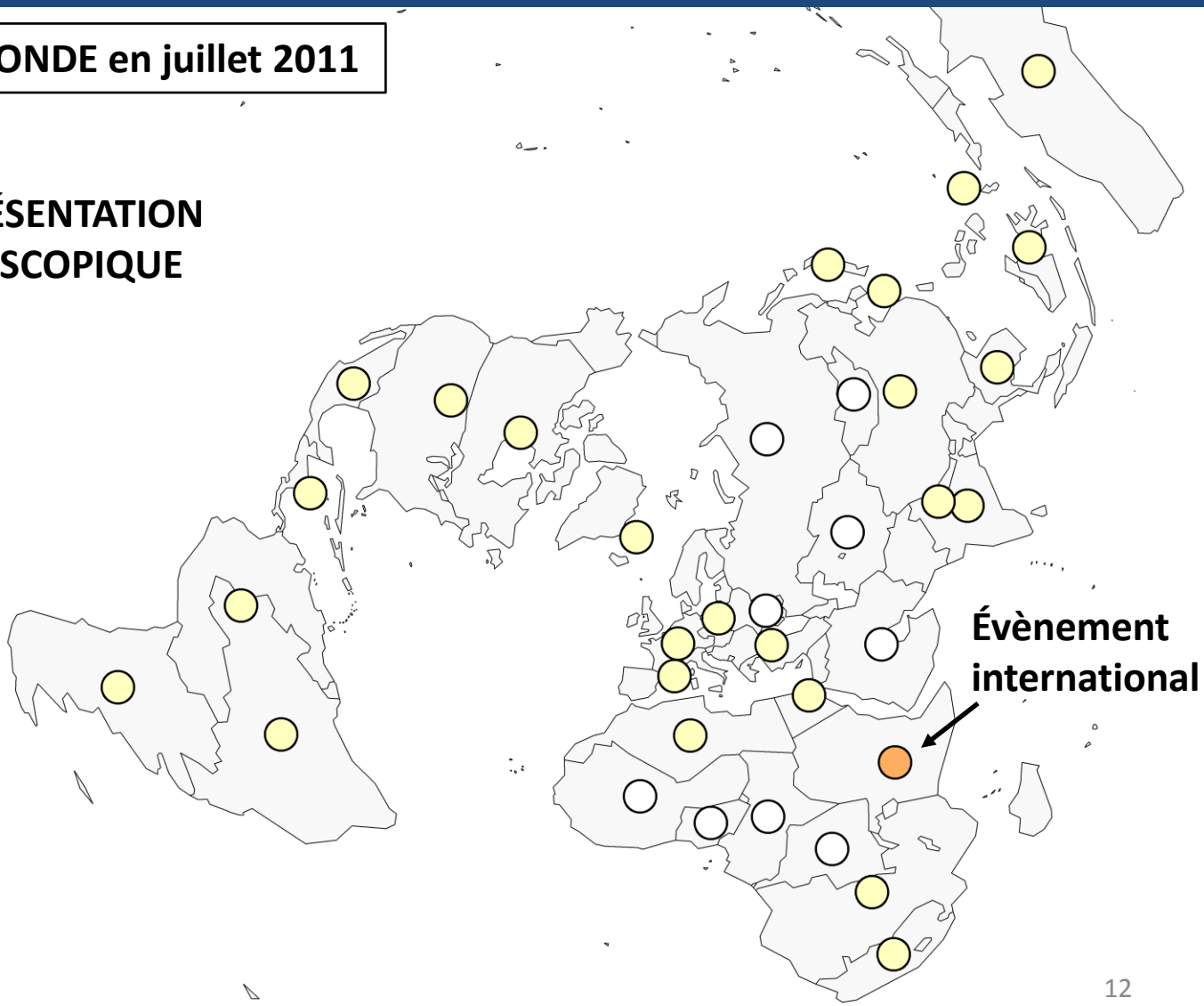
x 2.2

x 1.7

x 1.3

x 0.8

REPRÉSENTATION MÉSOSCOPIQUE



P2 : Niveaux de représentation

x 2.9

x 2.2

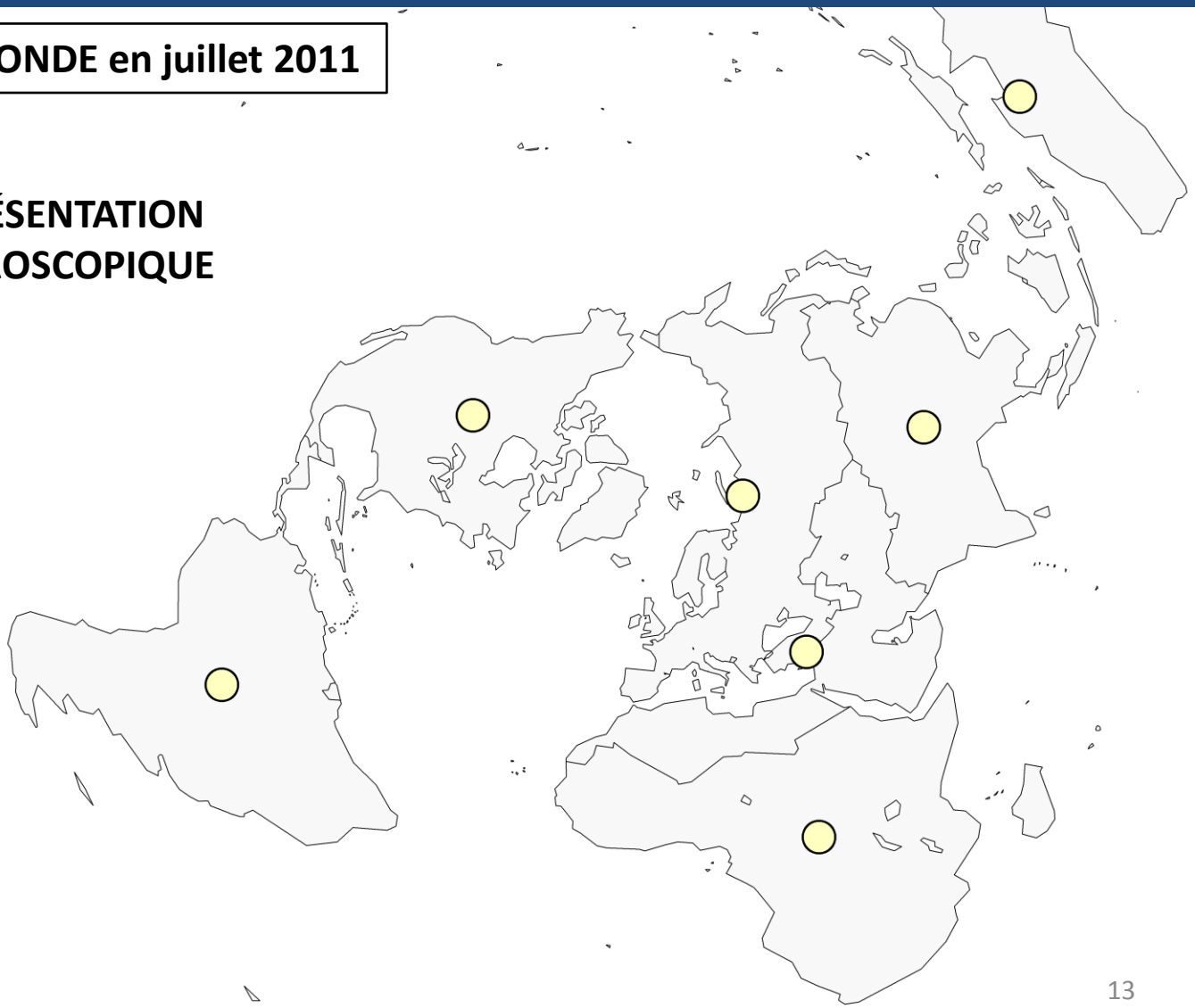
x 1.7

x 1.3

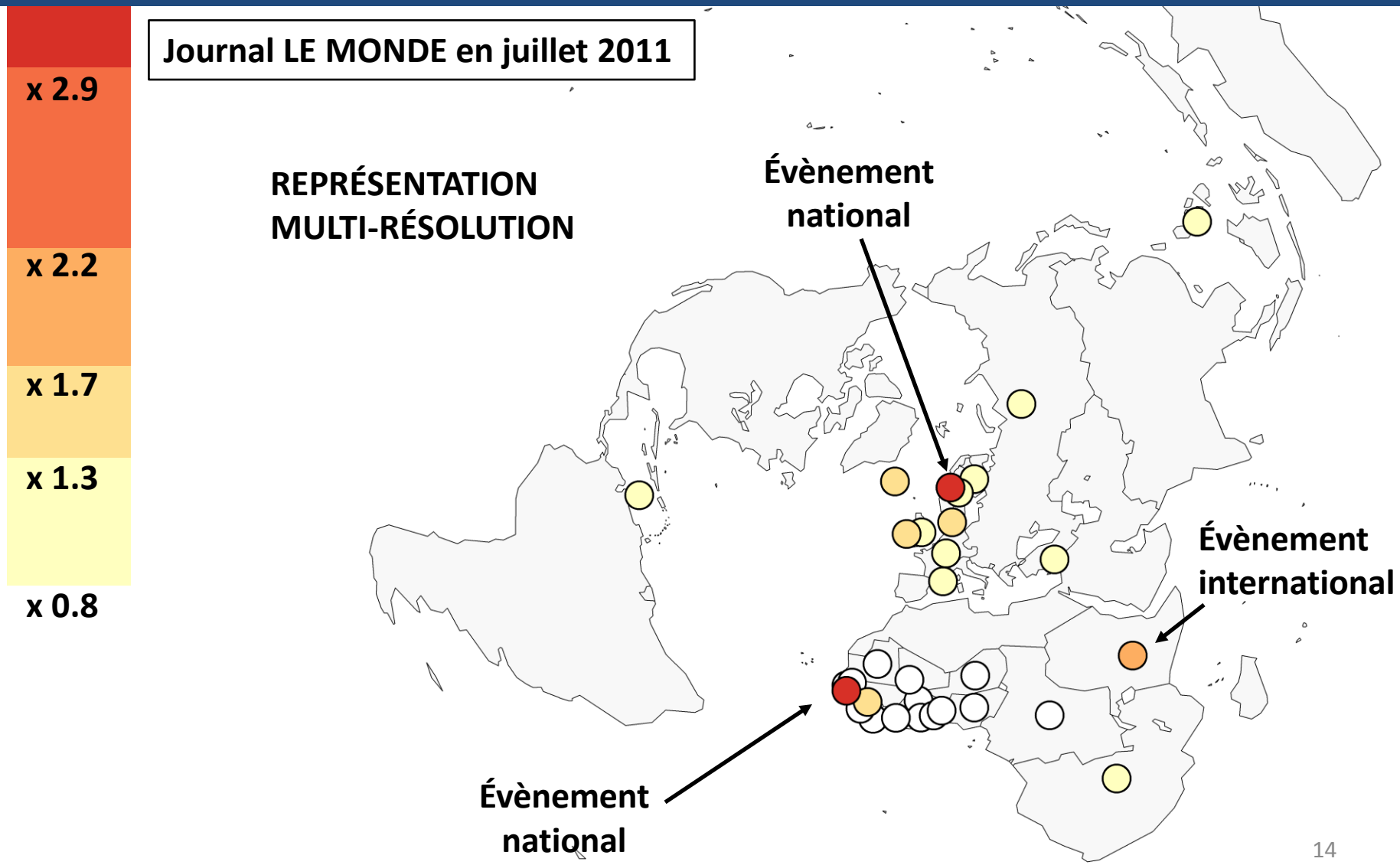
x 0.8

Journal LE MONDE en juillet 2011

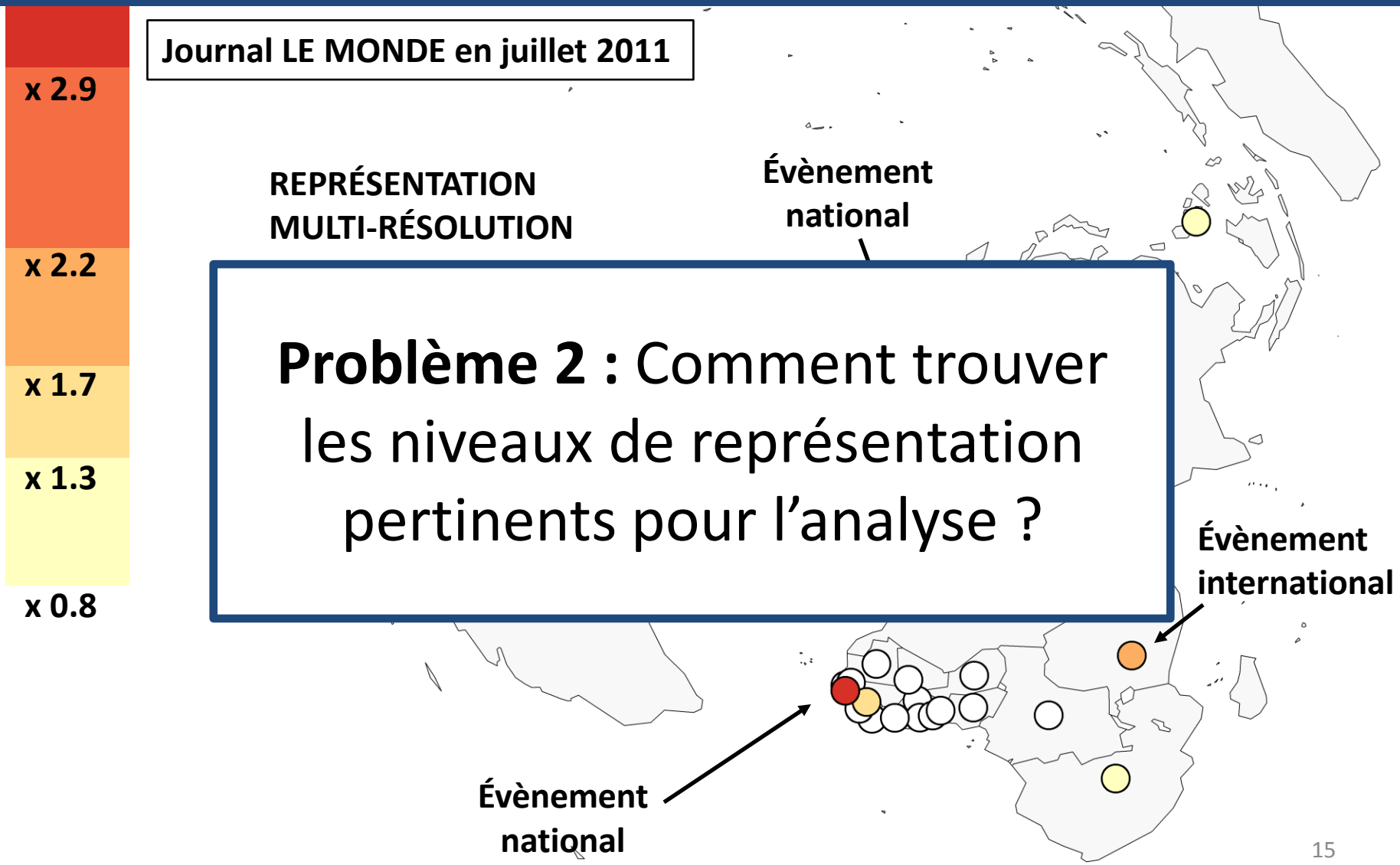
REPRÉSENTATION MACROSCOPIQUE



P2 : Niveaux de représentation

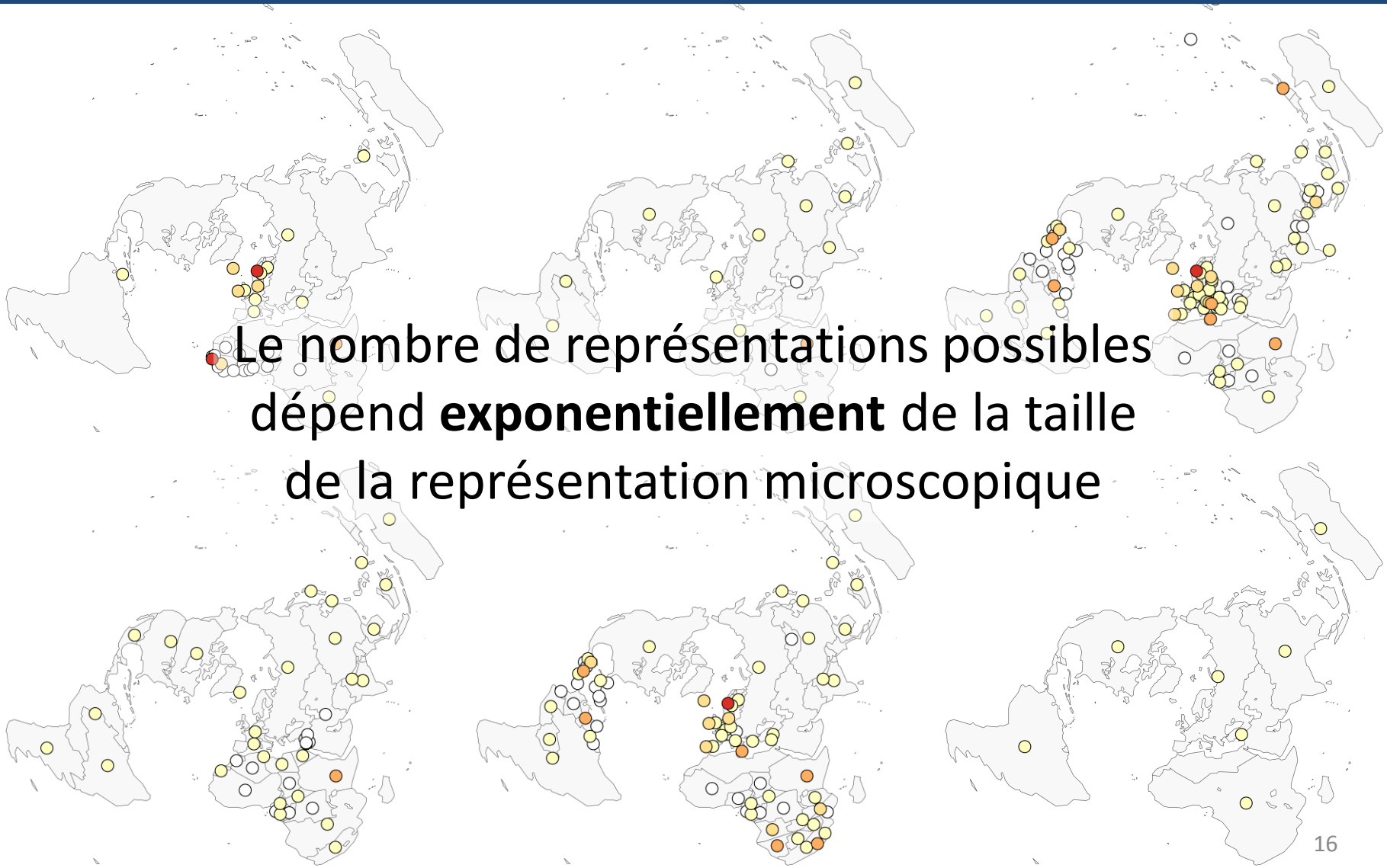


P2 : Niveaux de représentation



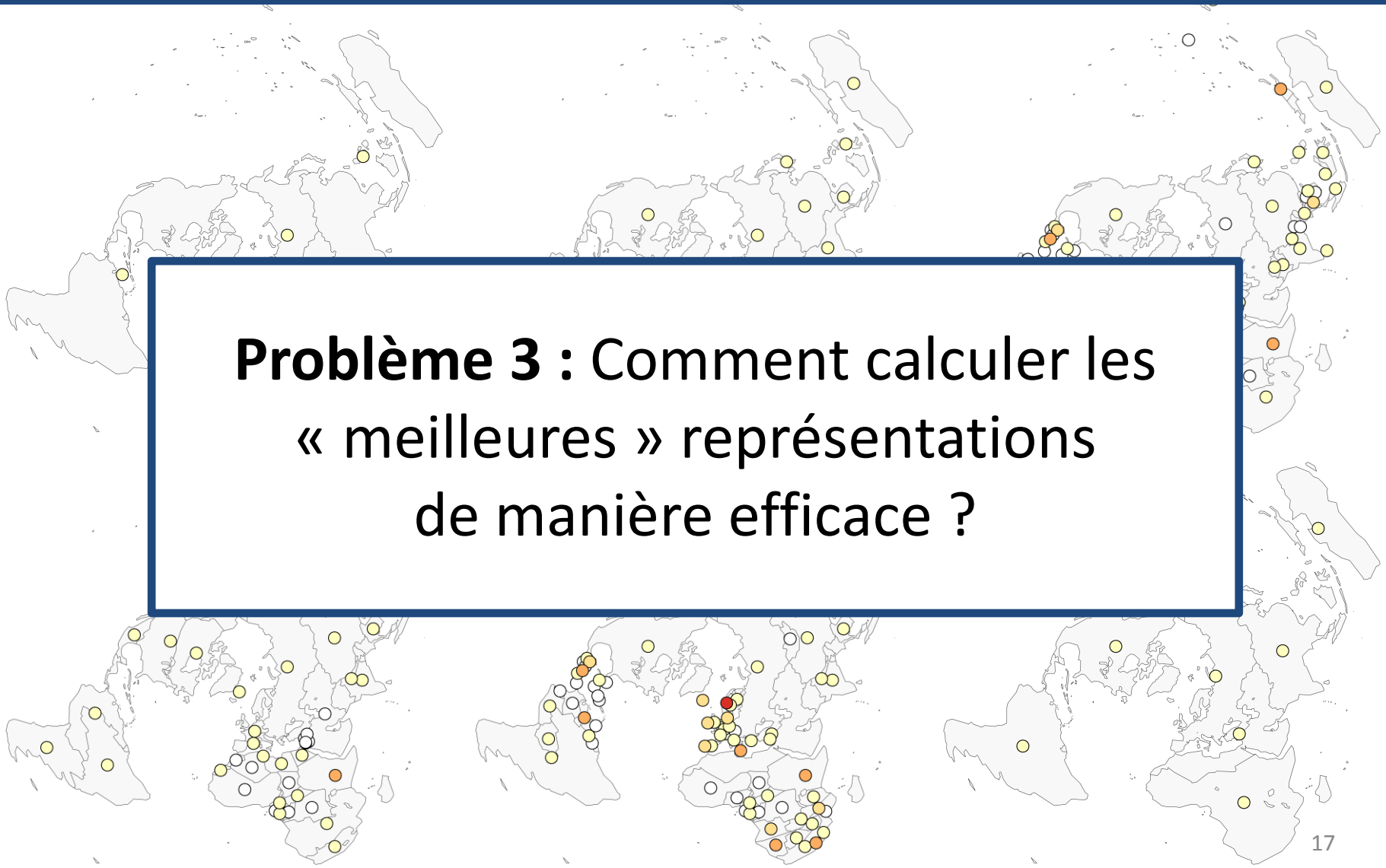
Problème 2 : Comment trouver les niveaux de représentation pertinents pour l'analyse ?

P3 : Calcul des meilleurs représentations



Le nombre de représentations possibles dépend **exponentiellement** de la taille de la représentation microscopique

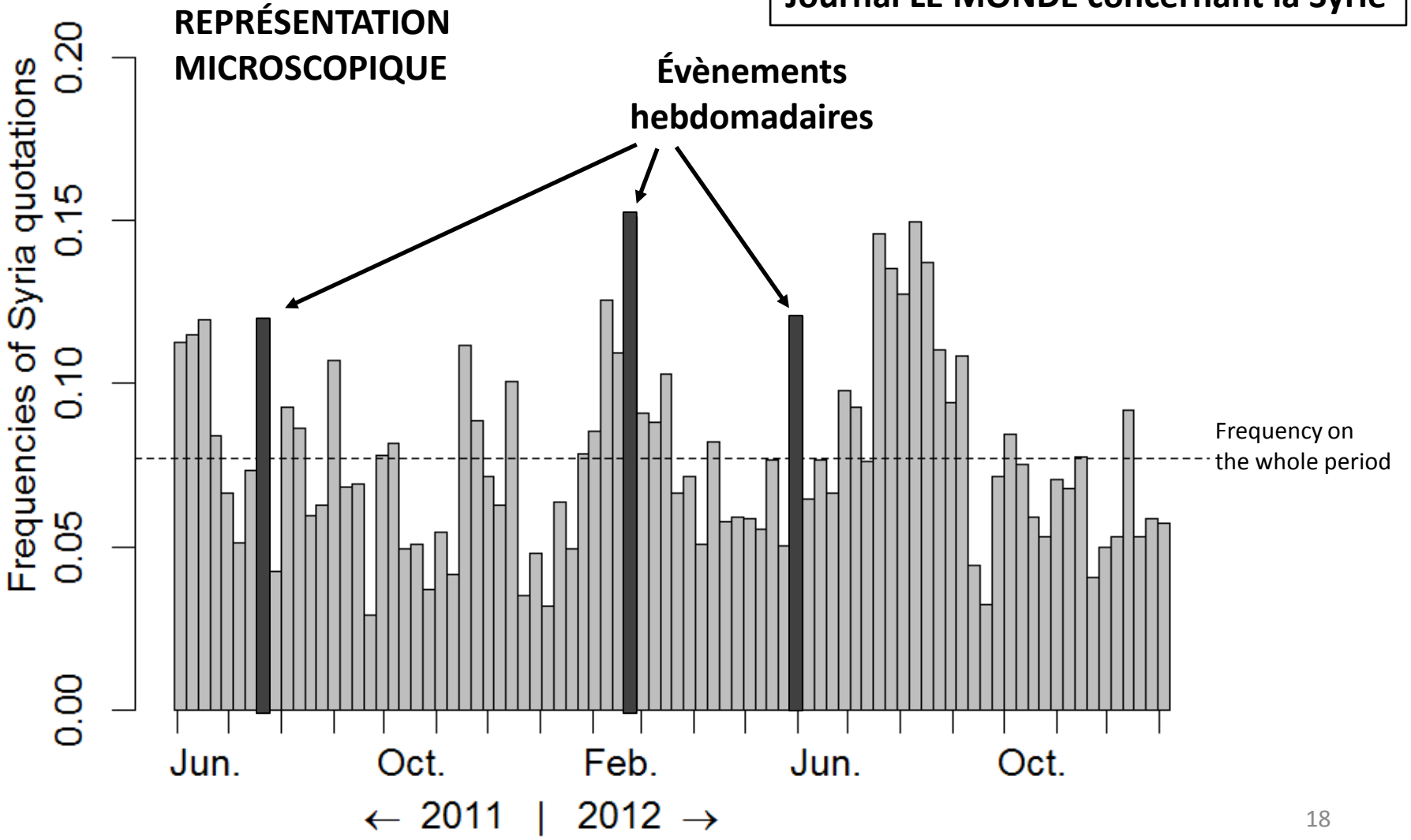
P3 : Calcul des meilleurs représentations



Problème 3 : Comment calculer les « meilleures » représentations de manière efficace ?

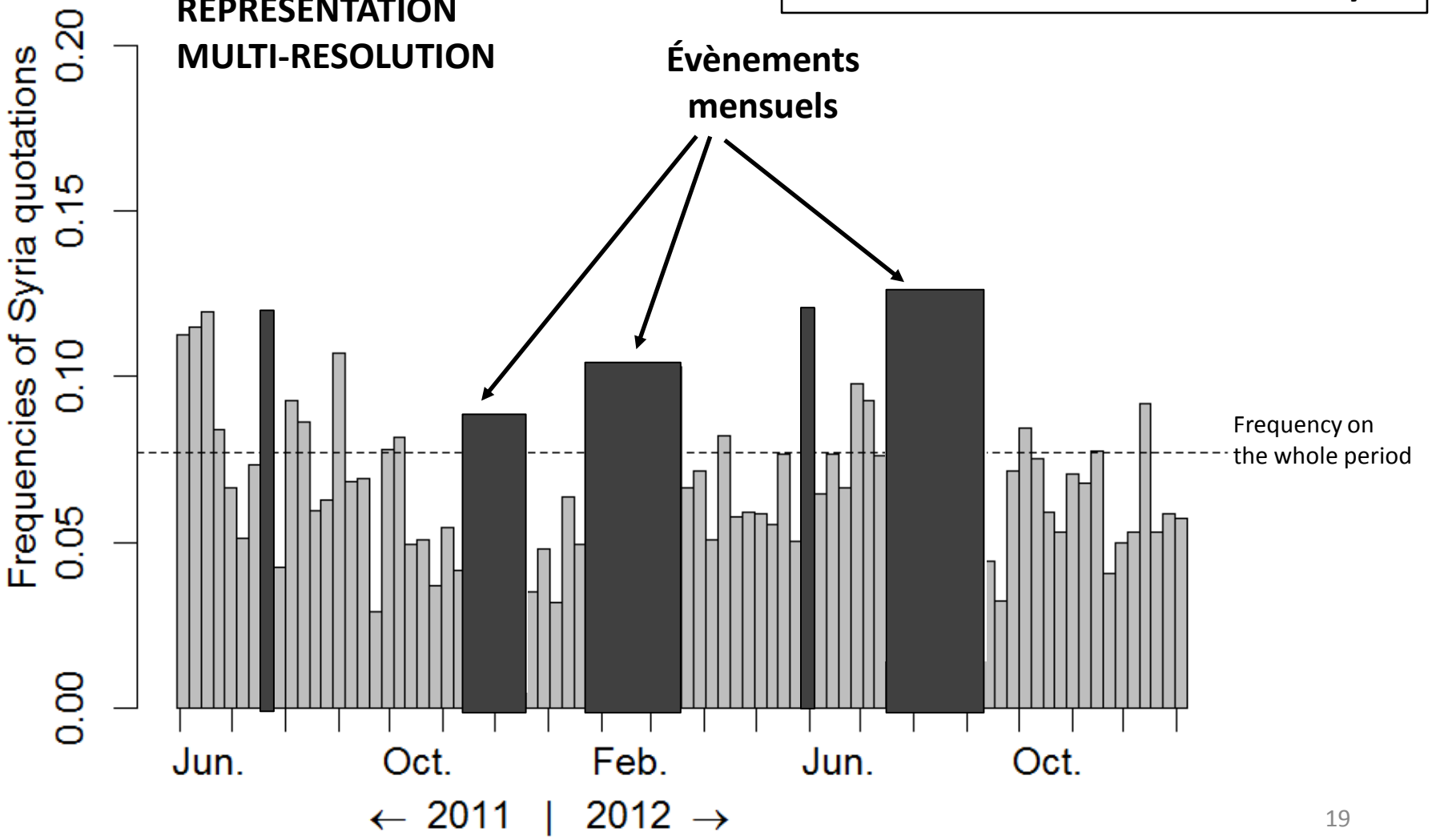
Et pour d'autres dimensions du système ?

Journal LE MONDE concernant la Syrie



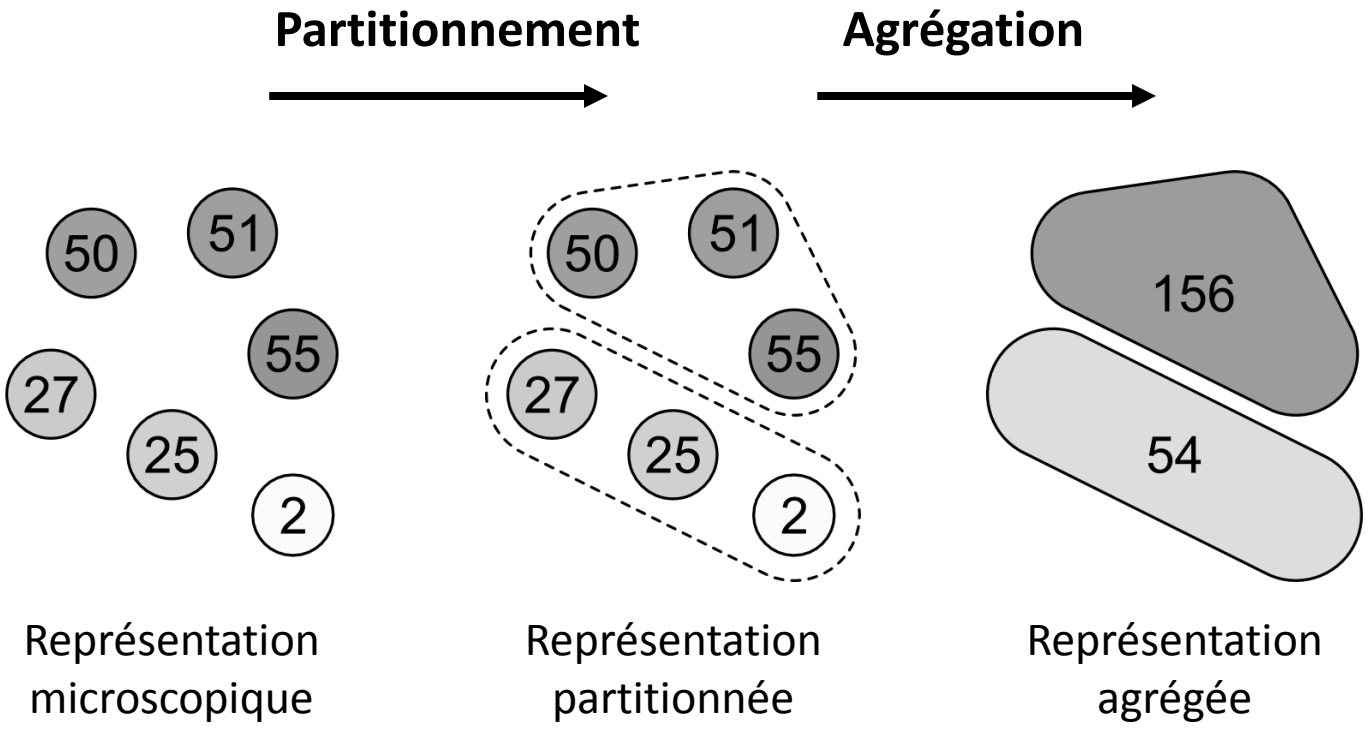
Et pour d'autres dimensions du système ?

Journal LE MONDE concernant la Syrie

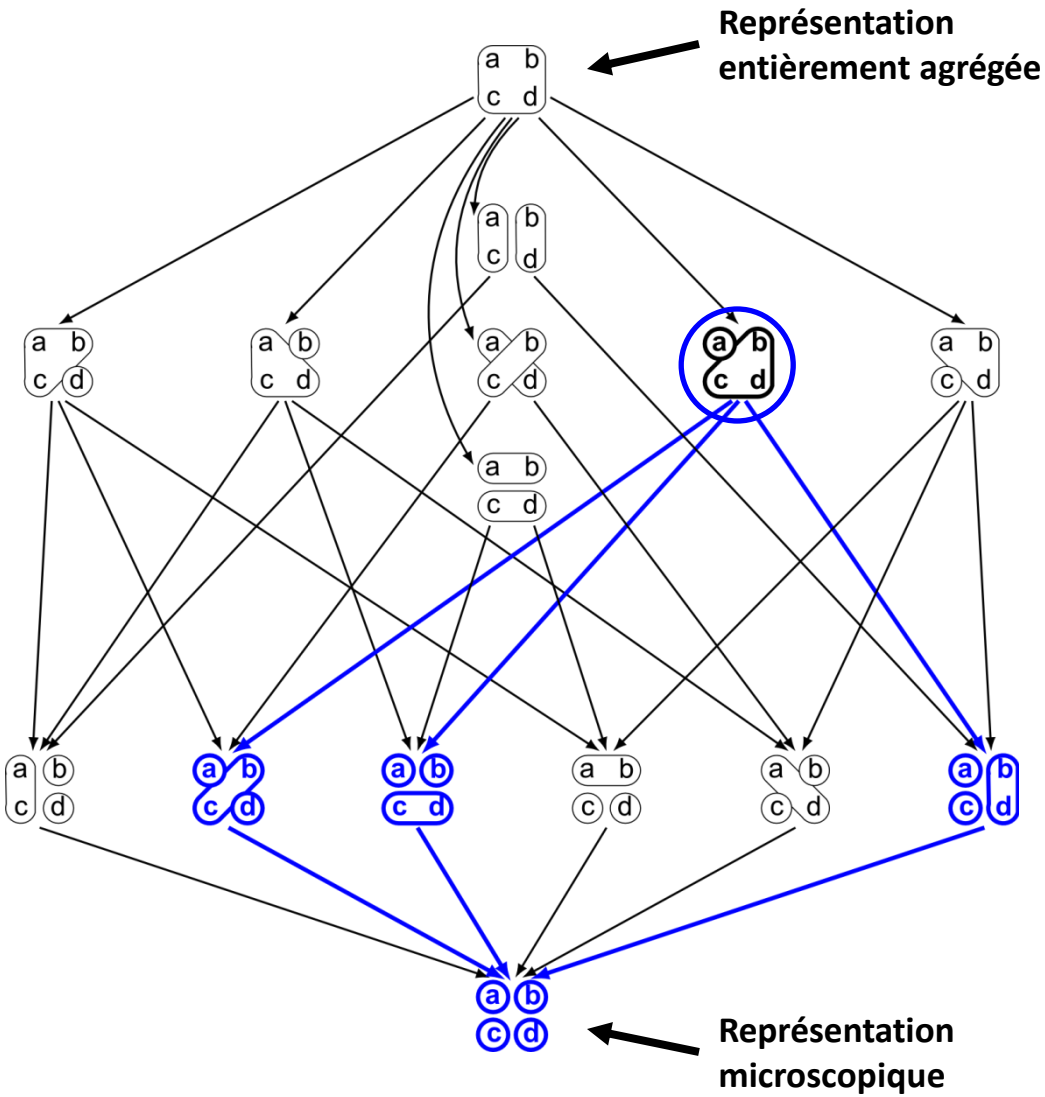


PRÉSERVER LA STRUCTURE DE L'ESPACE GÉOGRAPHIQUE

Le processus d'agrégation



Ensemble des partitions possibles



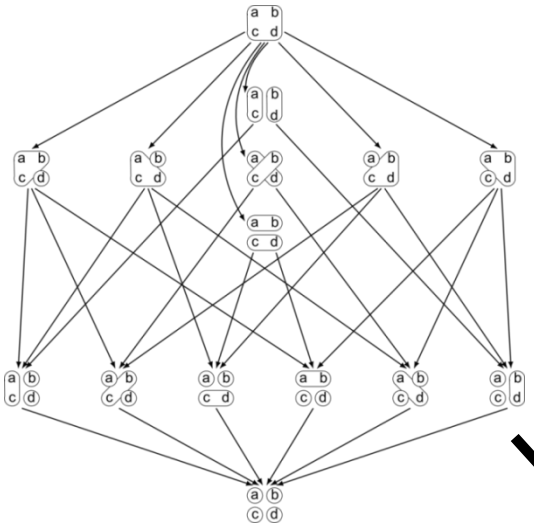
Structure algébrique

Ordre partiel sur l'ensemble des partitions possibles

→ **relations de raffinement**

Problème et objectif

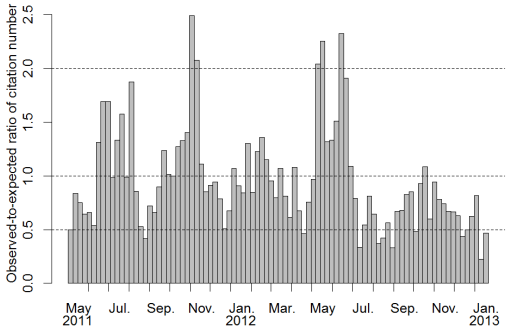
Ensemble des partitions <u>possibles</u>



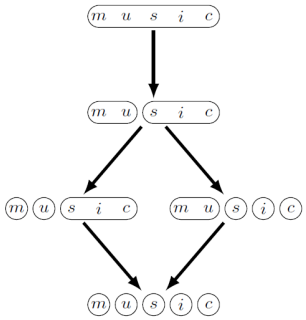
Sémantique géographique

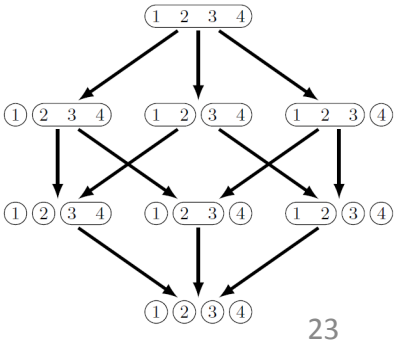
Géographe

Sémantique temporelle



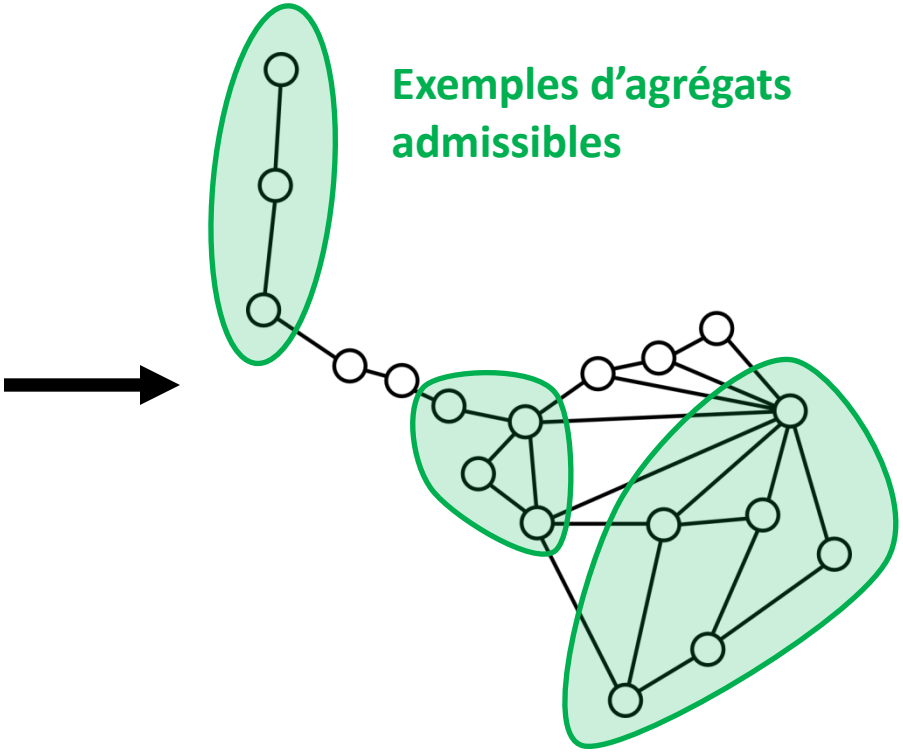
Contraindre l'ensemble des partitions <u>admissibles</u>





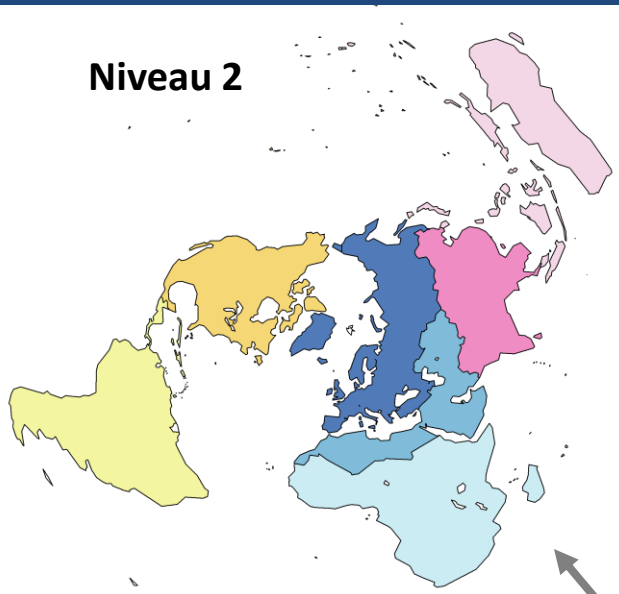
Conserver la relation de voisinage

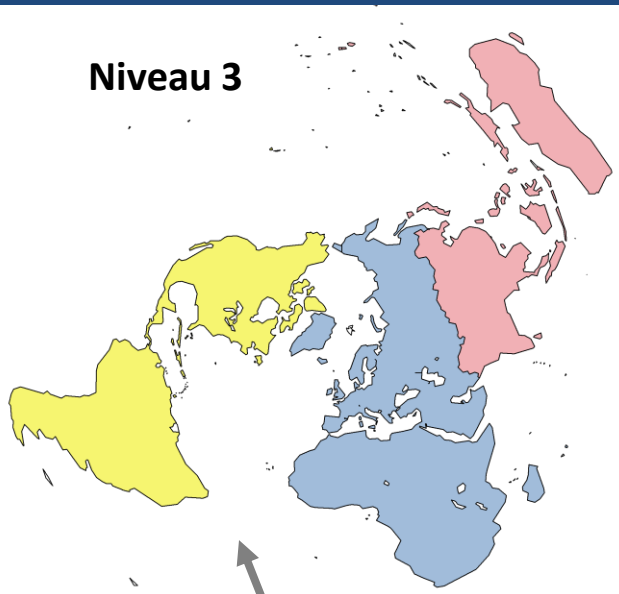
Agrégats admissibles : ensembles de pays connexes vis-à-vis du graphe de voisinage

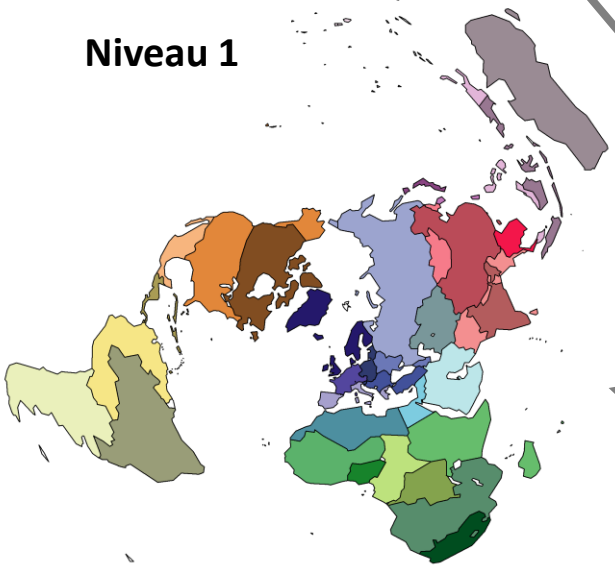


La hiérarchie WUTS

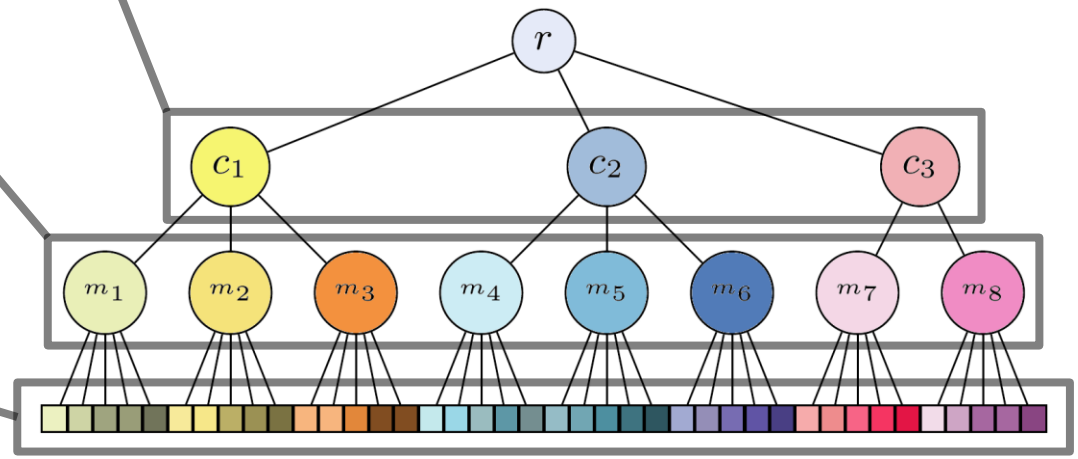
[Grasland et Didelon, 2007]





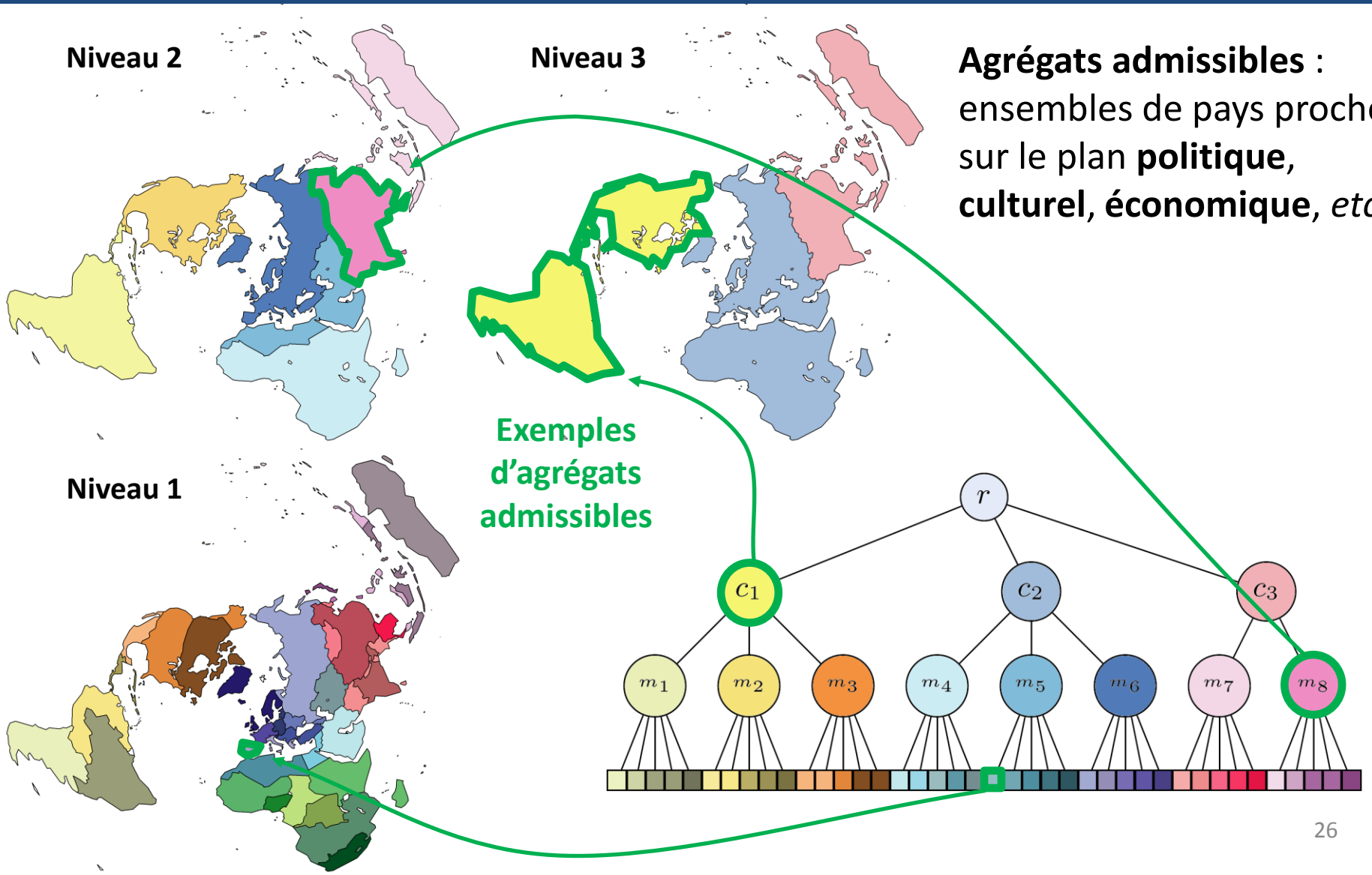


Agrégats admissibles : ensembles de pays proches sur le plan **politique**, **culturel**, **économique**, *etc.*



La hiérarchie WUTS

[Grasland et Didelon, 2007]

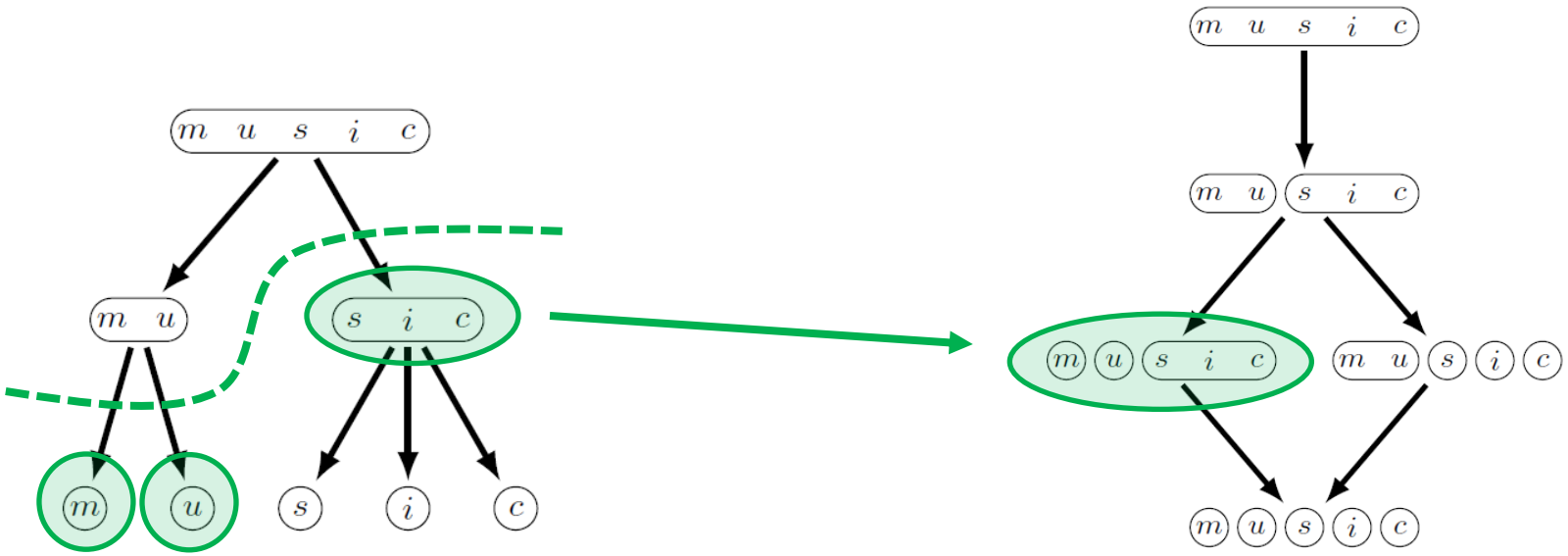


Agrégats admissibles : ensembles de pays proches sur le plan **politique**, **culturel**, **économique**, *etc.*

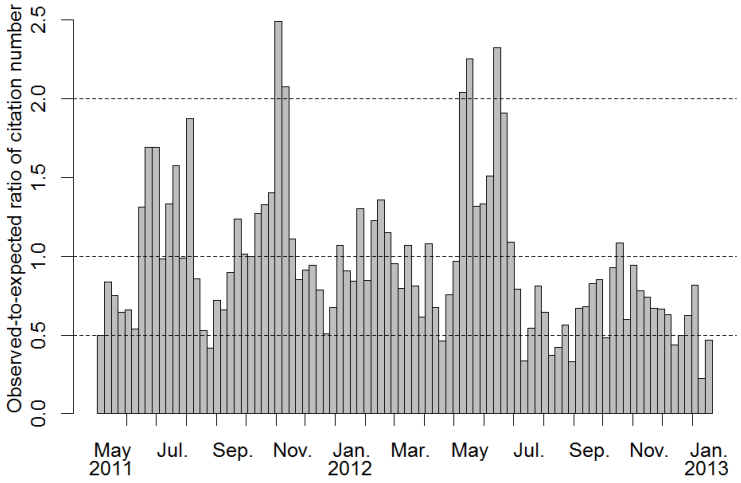
Agrégation selon une hiérarchie

Agrégats admissibles
(nœuds de la hiérarchie)

Partitions admissibles
(coupe dans la hiérarchie)

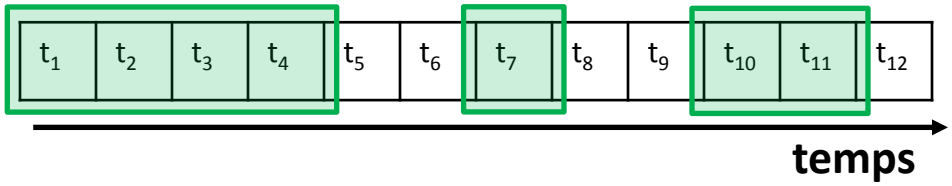


Conserver l'ordre du temps



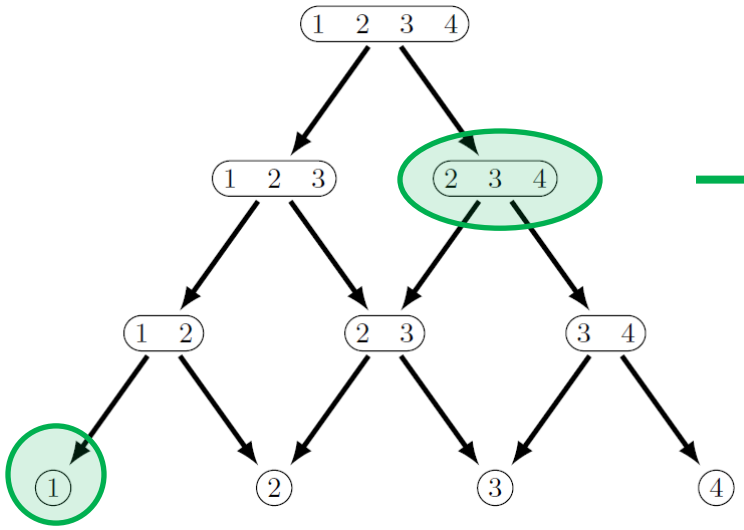
Agrégats admissibles :
intervalles de temps

Exemples d'agrégats admissibles

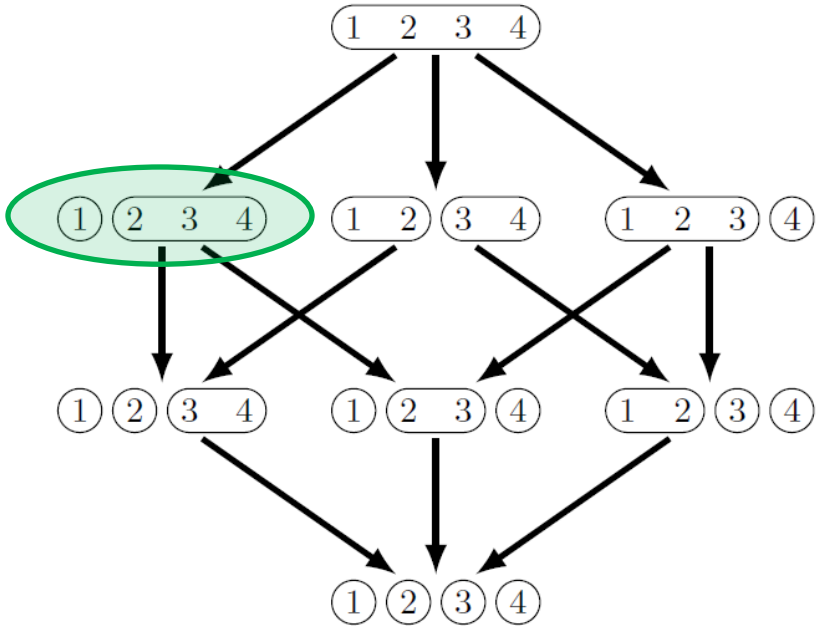


Agrégation selon un ordre total

Agrégats admissibles
(intervalles de temps)

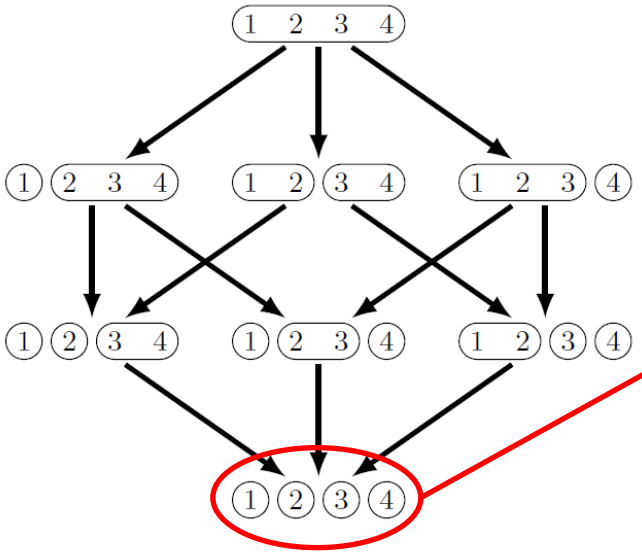


Partitions admissibles
(séquences d'intervalles)



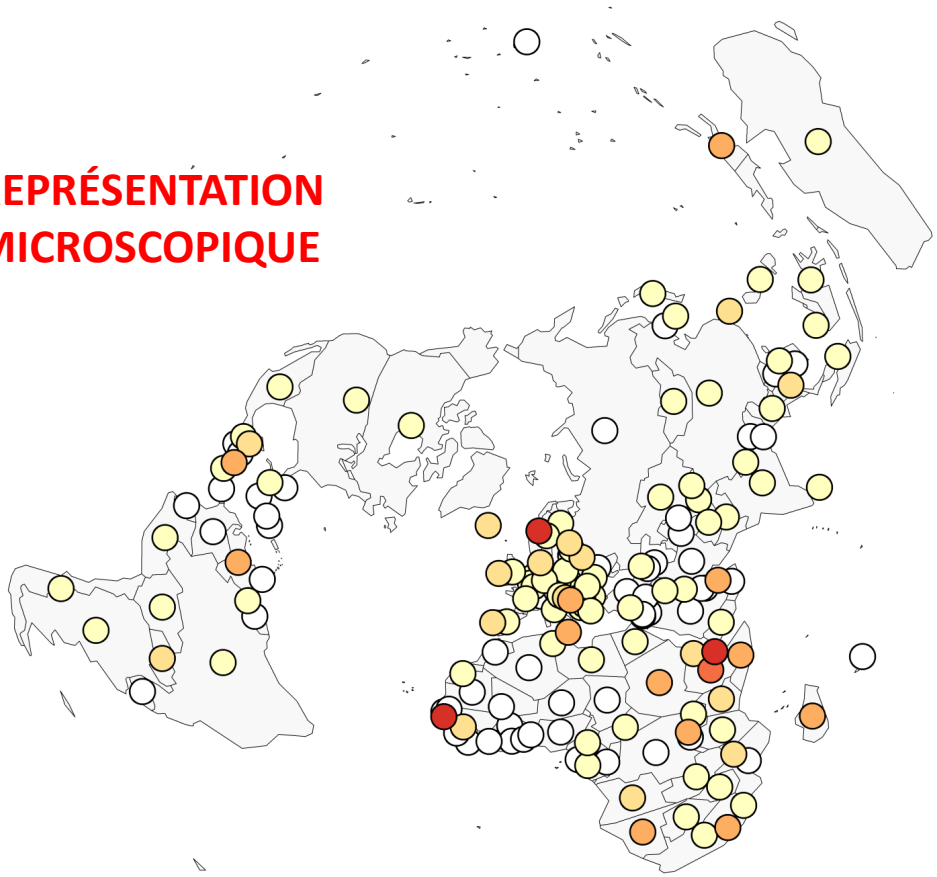
AGRÉGATION ET INFORMATION

Objectifs et difficultés



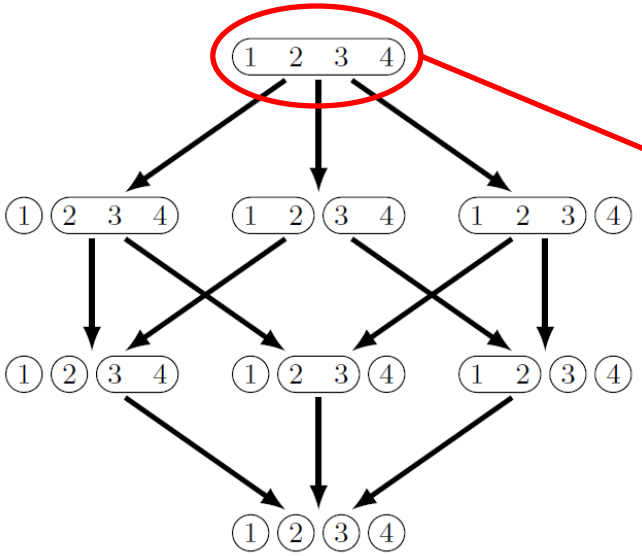
Quelle partition admissible est la meilleure pour un jeu de données particulier ?

REPRÉSENTATION MICROSCOPIQUE



→ TROP COMPLEXE POUR PASSER À L'ÉCHELLE

Objectifs et difficultés

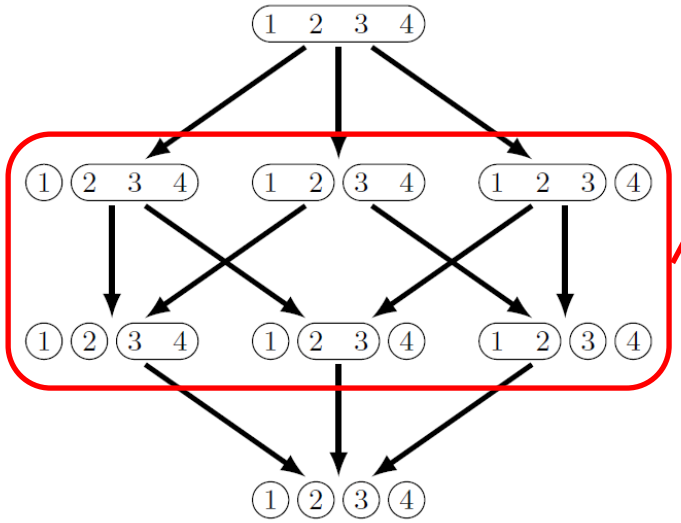


Quelle partition admissible est la meilleure pour un jeu de données particulier ?

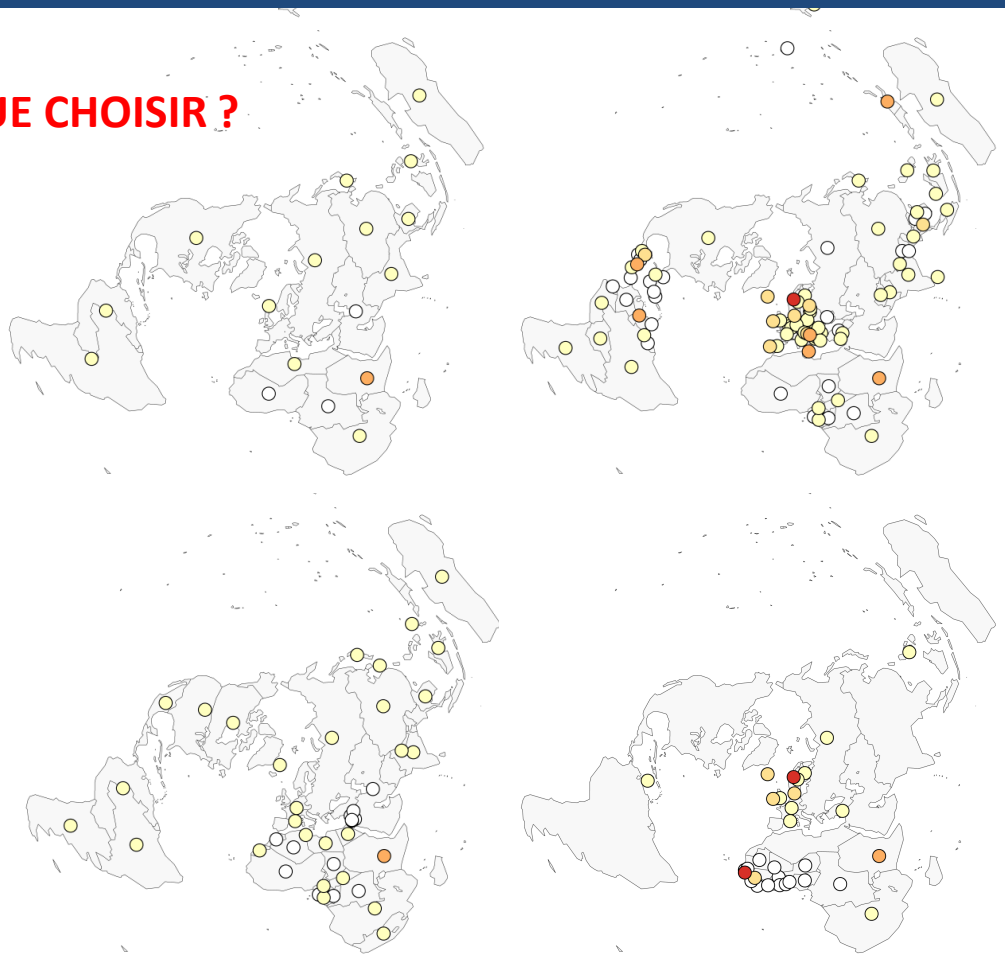
REPRÉSENTATION ENTIÈREMENT AGRÉGÉE

→ NE DONNE QUE TRÈS PEU D'INFORMATION

Objectifs et difficultés

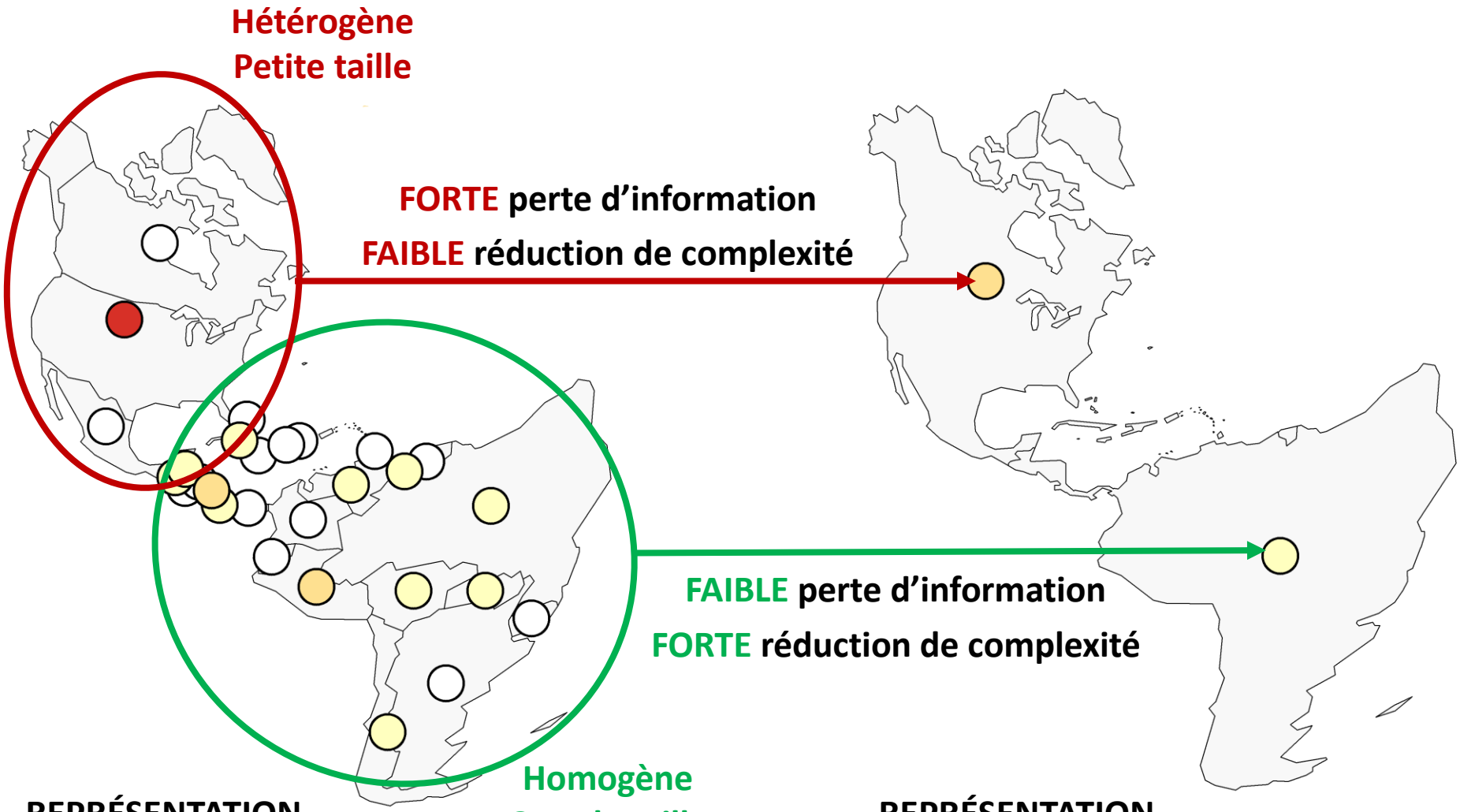


QUE CHOISIR ?

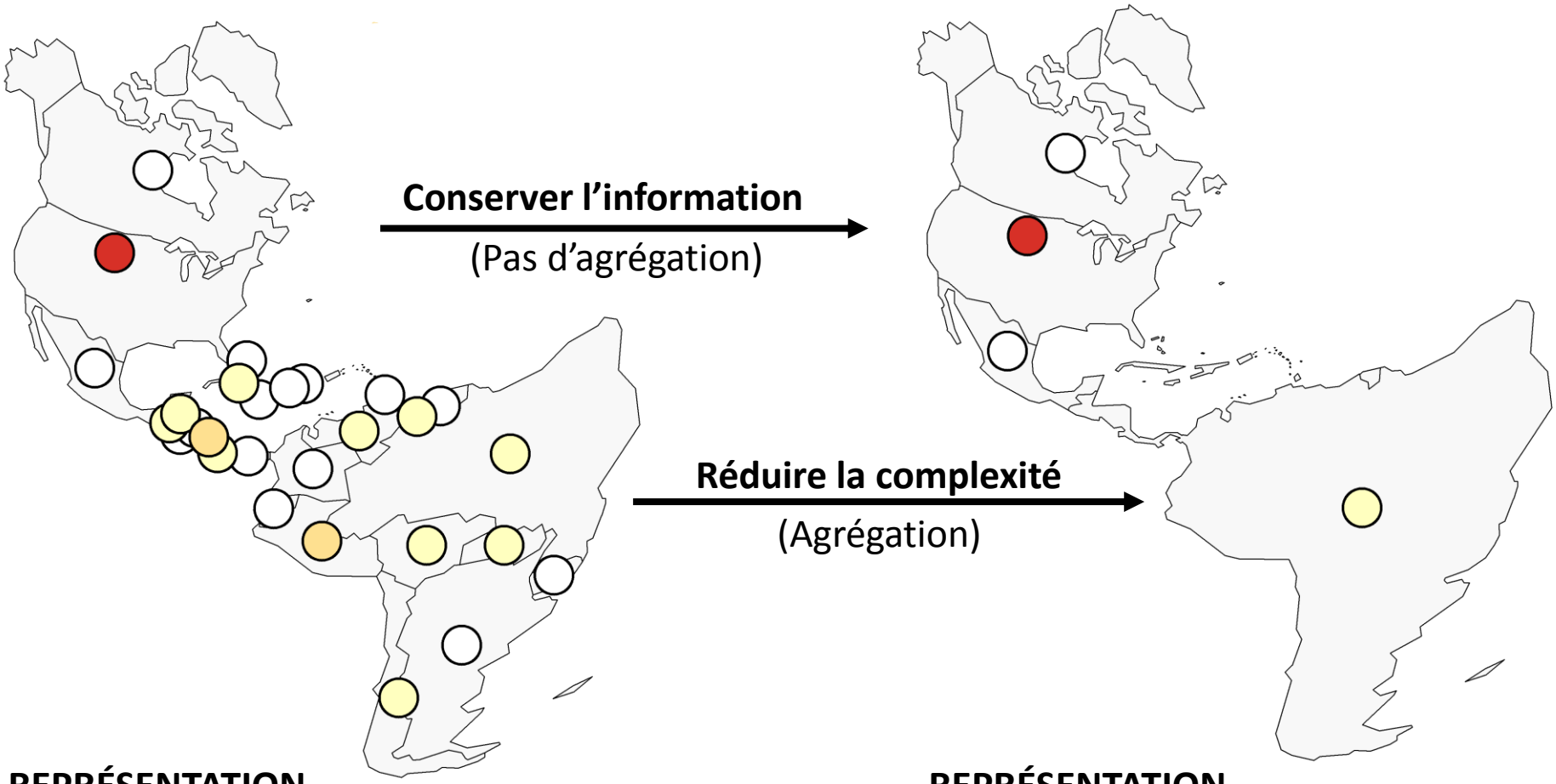


Quelle partition admissible est la meilleure pour un jeu de données particulier ?

Complexité et information



Complexité et information



REPRÉSENTATION MICROSCOPIQUE

REPRÉSENTATION MULTI-RÉSOLUTION

Mesures de qualité

[Lamarche-Perrin *et al.*, ECCS 2012]

La **complexité** dépend de la **tâche à accomplir** et des **outils de description** disponibles

[Bonabeau et Dessalles, 1997]

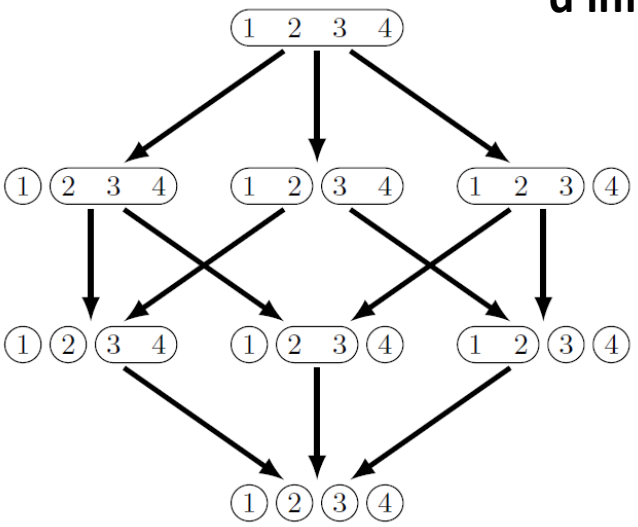
La **perte d'information** est mesurée par la **divergence** entre deux distributions de probabilité

[Kullback et Leibler, 1951]

Perte d'information

Nombre d'agrégats représentés :

$$T(\mathcal{X}) = |\mathcal{X}|$$



Divergence de Kullback-Leibler :

$$D(\mathcal{X}) = \sum_{X \in \mathcal{X}} \sum_{x \in X} \frac{v(x)}{v(\Omega)} \log_2 \left(\frac{v(x)\,|X|}{v(X)} \right)$$

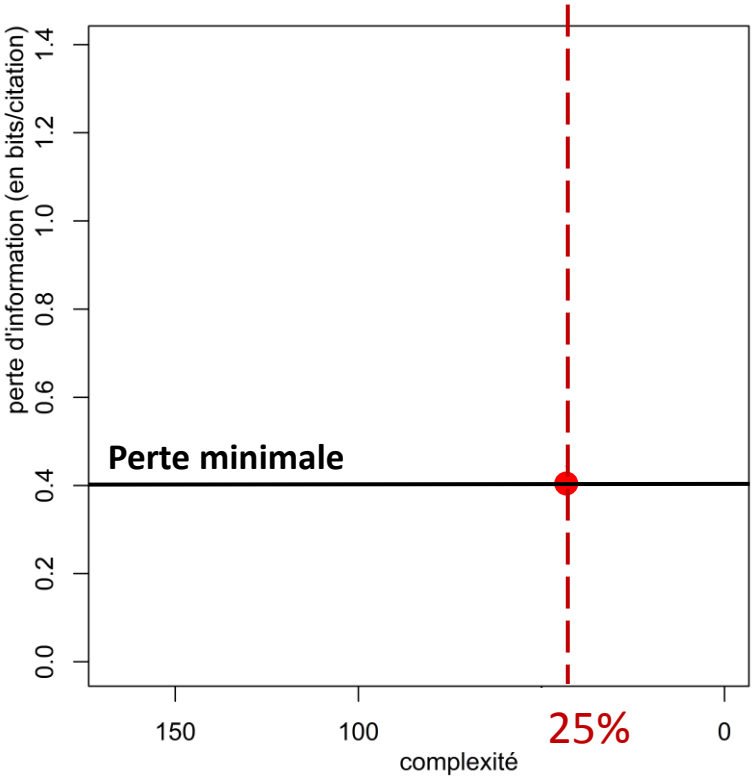
Complexité

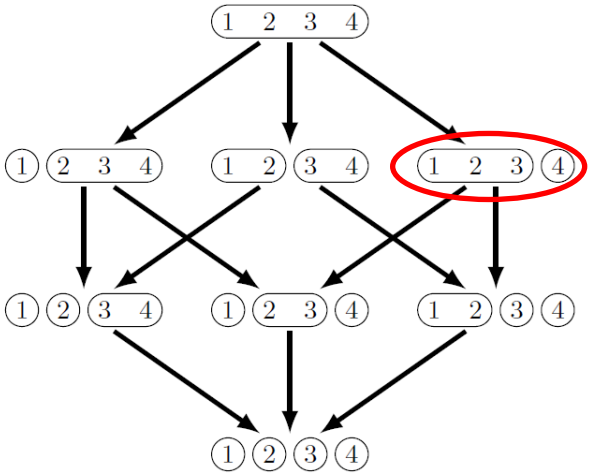
Optimisation des mesures de qualité

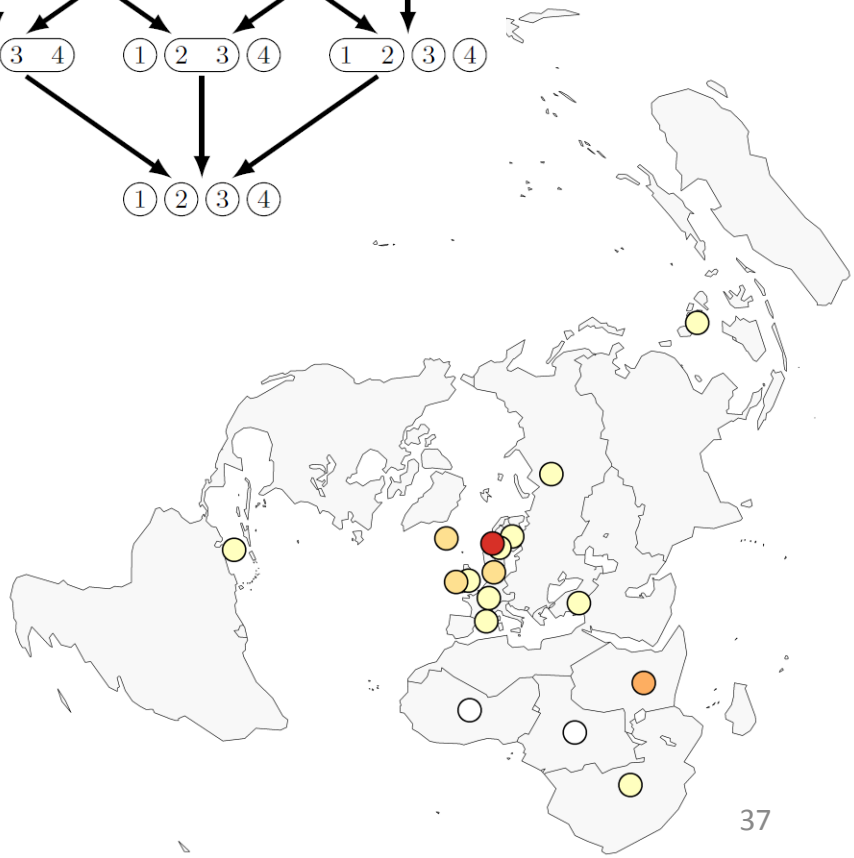
Deux critères d'évaluation indépendants

Compromis de qualité :

$$CQL_\alpha = \alpha \frac{\Delta T}{\Delta T_{\max}} - (1-\alpha)\frac{D}{D_{\max}}$$





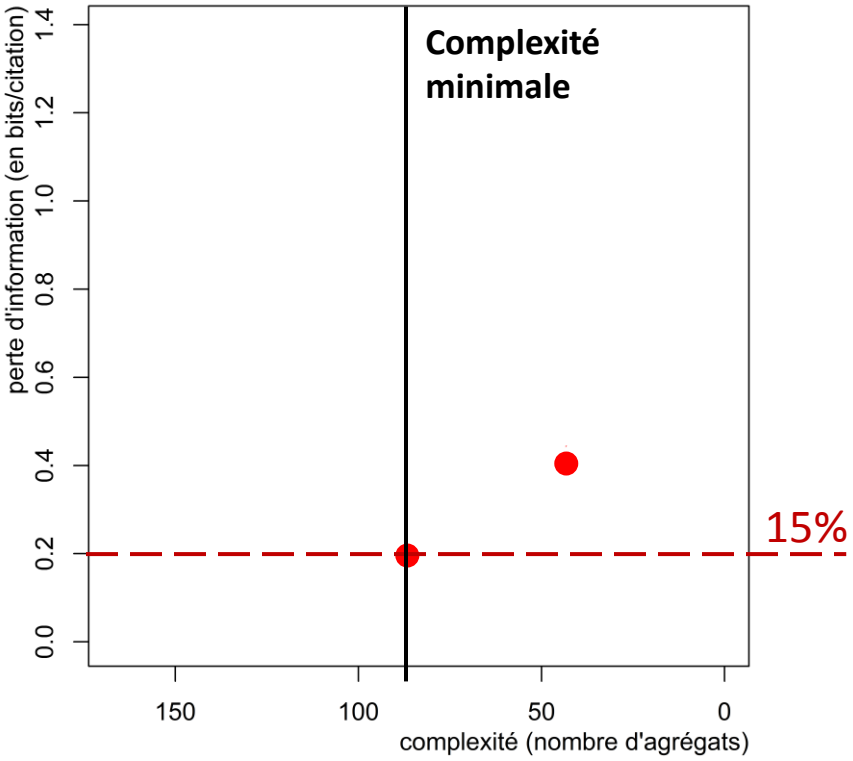


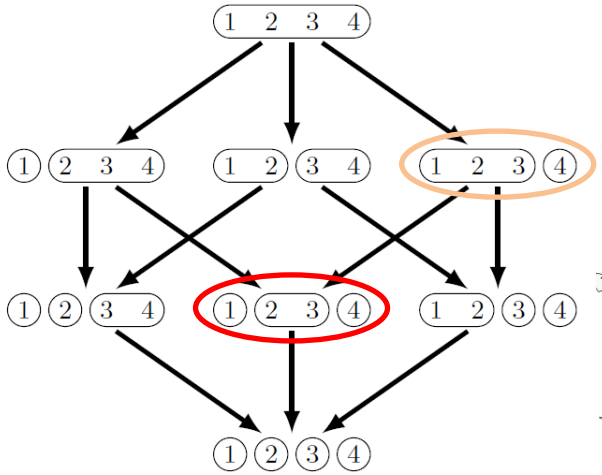
Optimisation des mesures de qualité

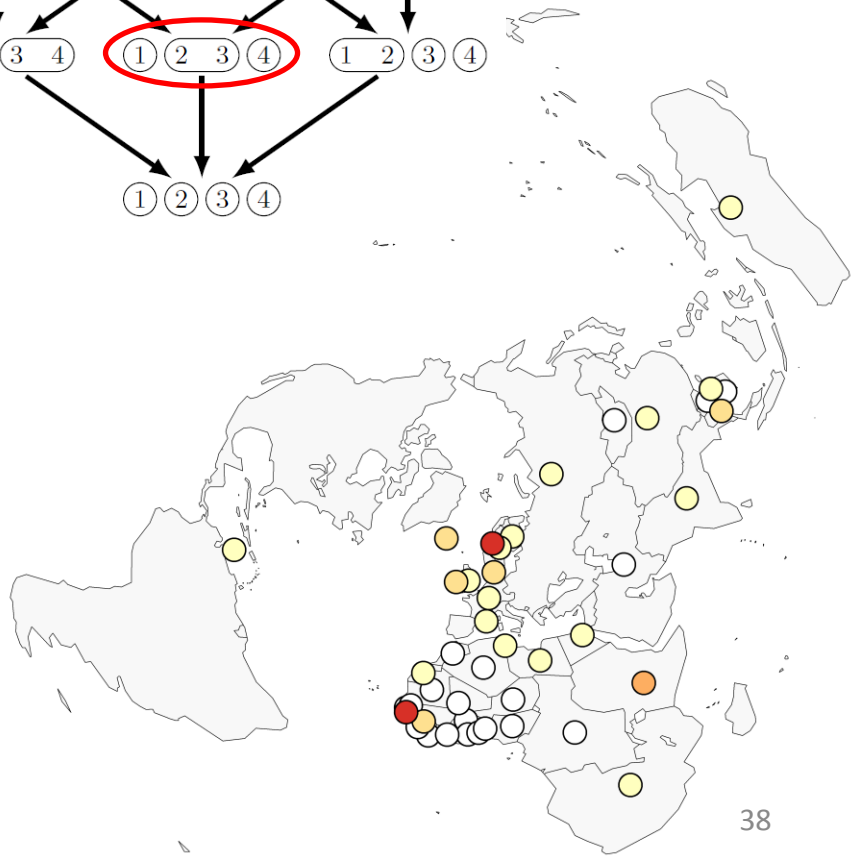
Deux critères d'évaluation indépendants

Compromis de qualité :

$$CQL_\alpha = \alpha \frac{\Delta T}{\Delta T_{\max}} - (1-\alpha)\frac{D}{D_{\max}}$$





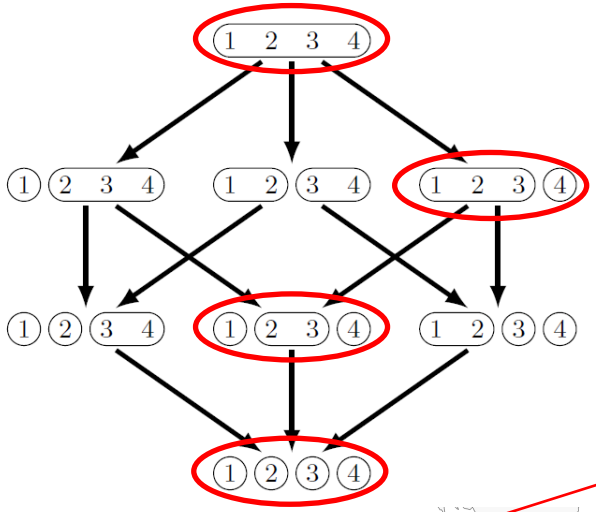


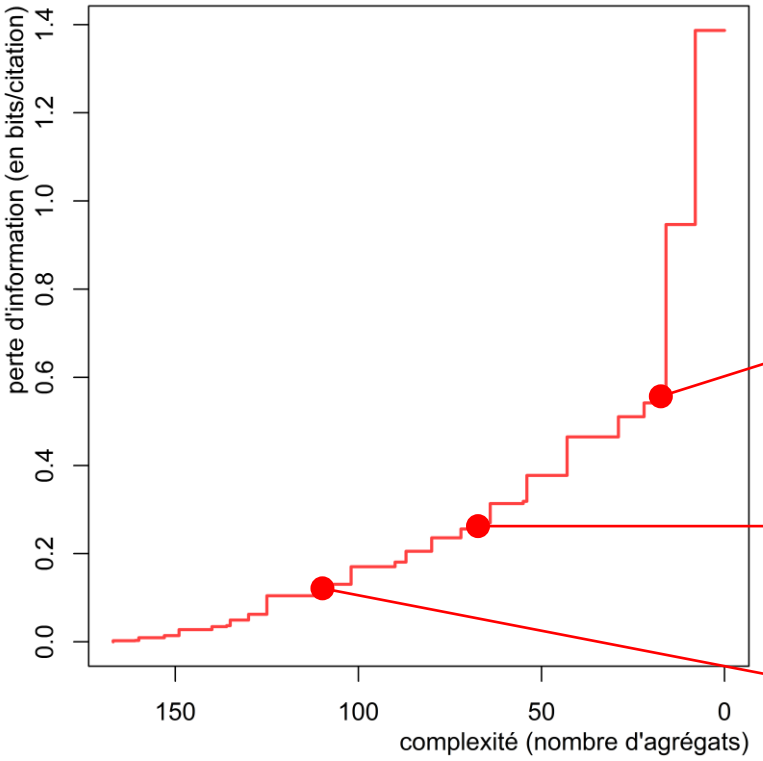
Optimisation des mesures de qualité

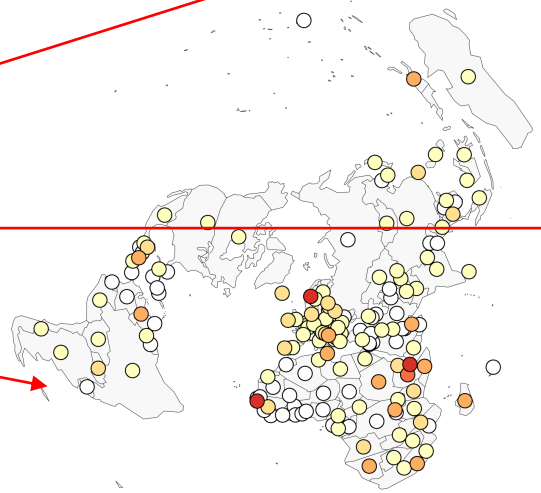
Deux critères d'évaluation indépendants

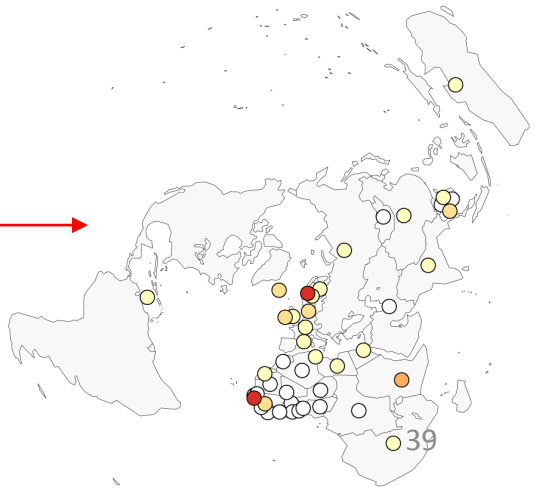
Compromis de qualité :

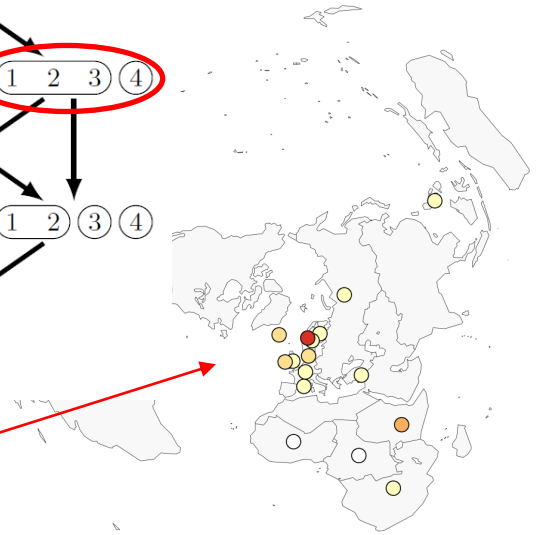
$$CQL_\alpha = \alpha \frac{\Delta T}{\Delta T_{\max}} - (1-\alpha)\frac{D}{D_{\max}}$$





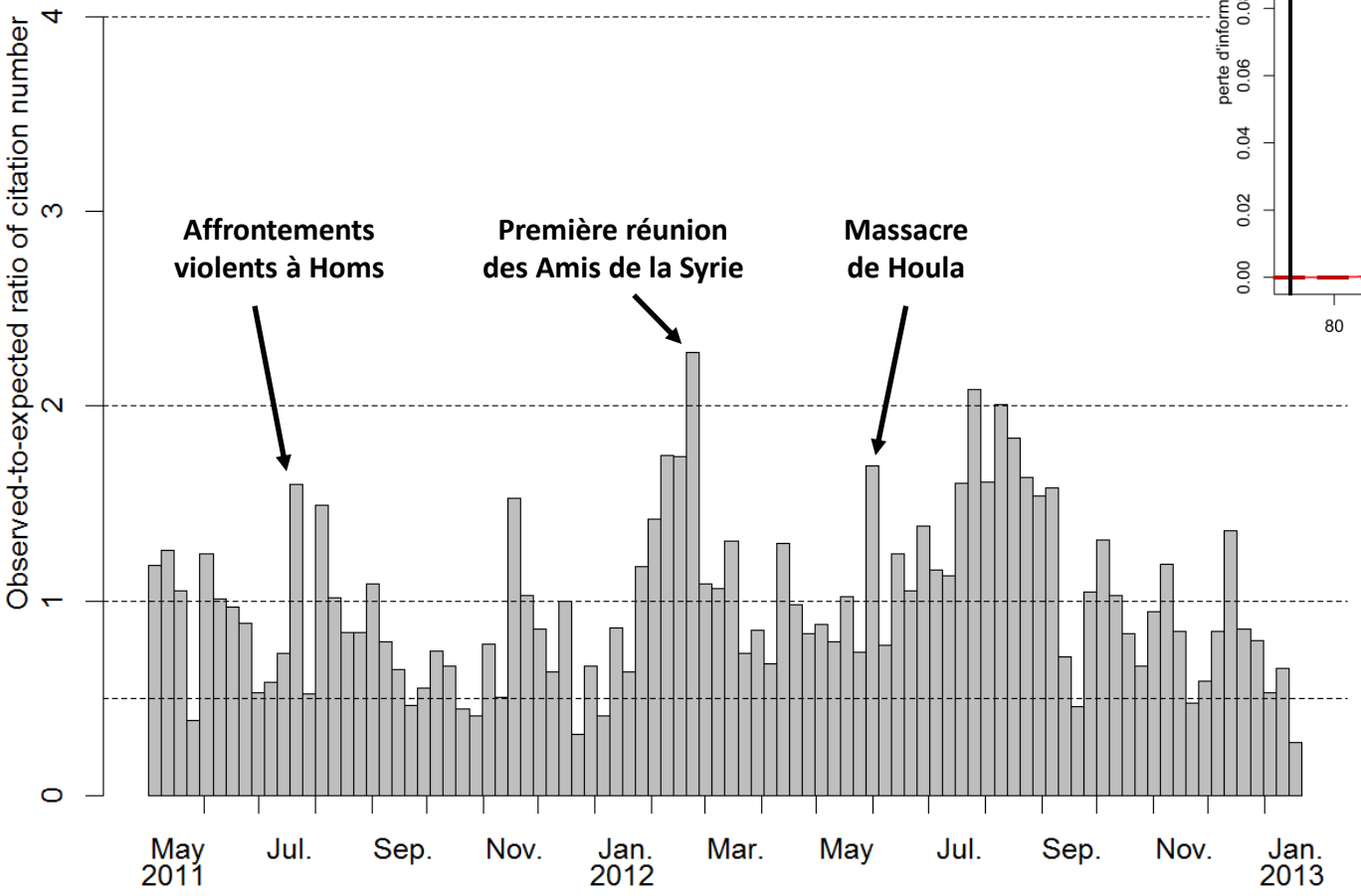


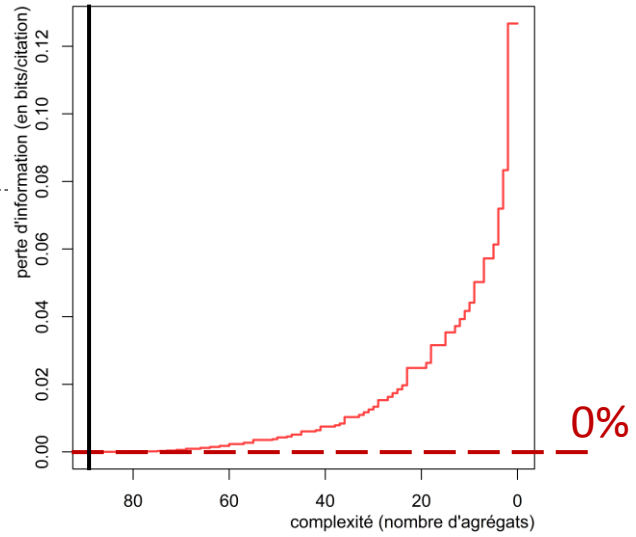




La Syrie vue par LE MONDE

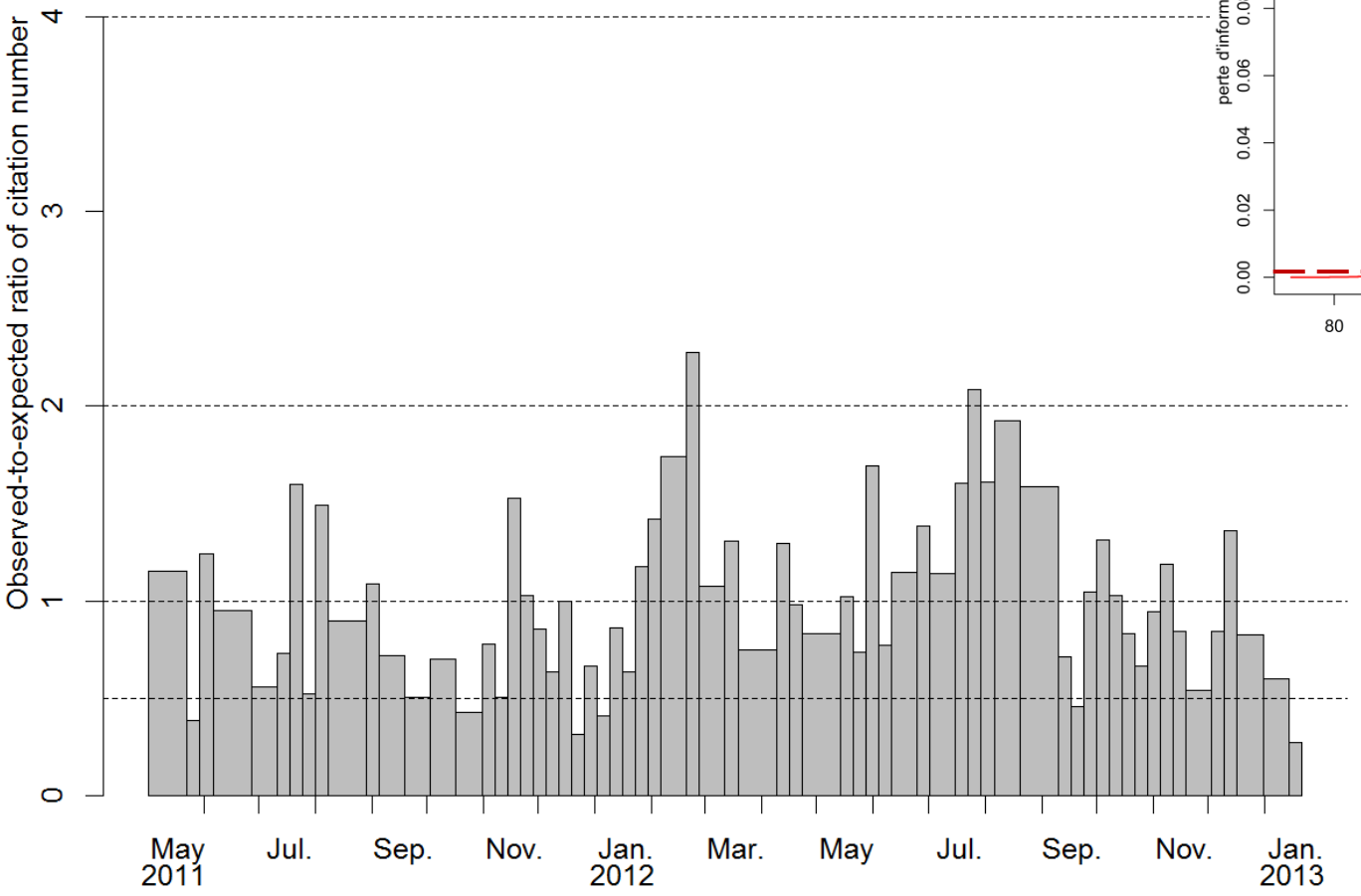
[Giraud, Grasland, Lamarche-Perrin *et al.*, ECTQG 2013]

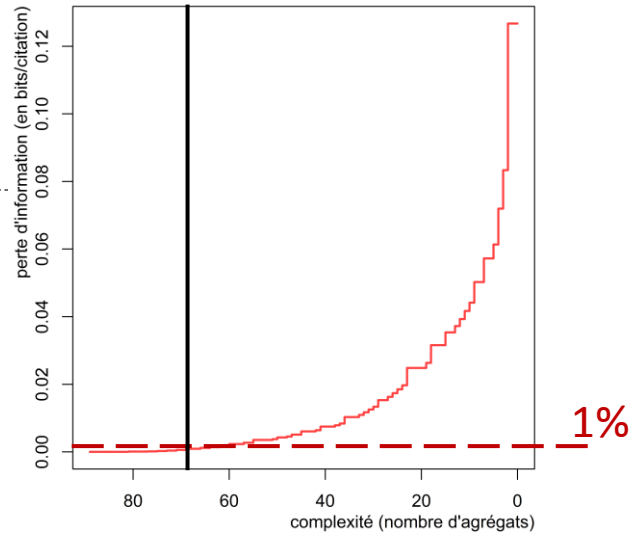




La Syrie vue par LE MONDE

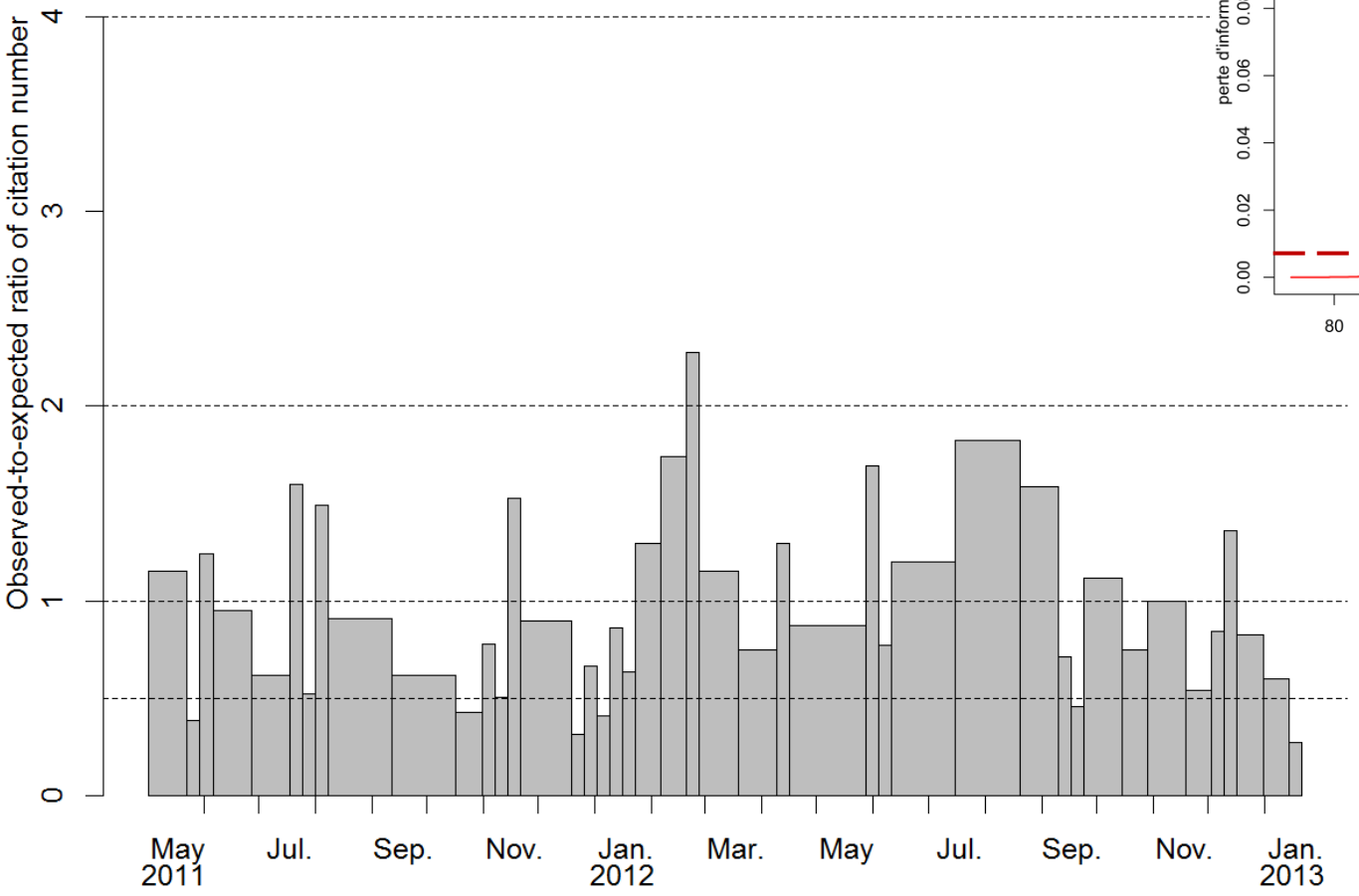
[Giraud, Grasland, Lamarche-Perrin *et al.*, ECTQG 2013]

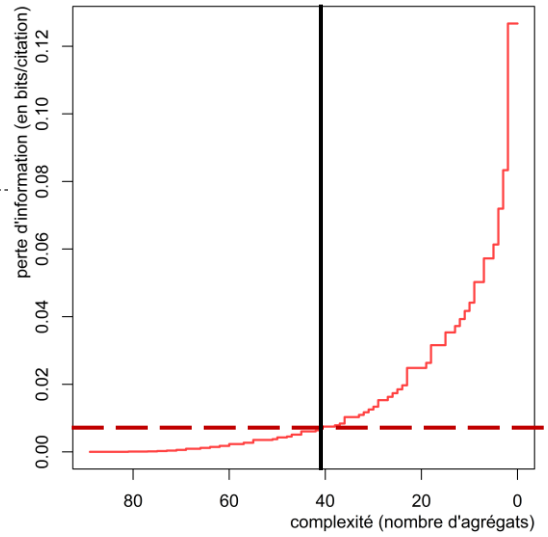




La Syrie vue par LE MONDE

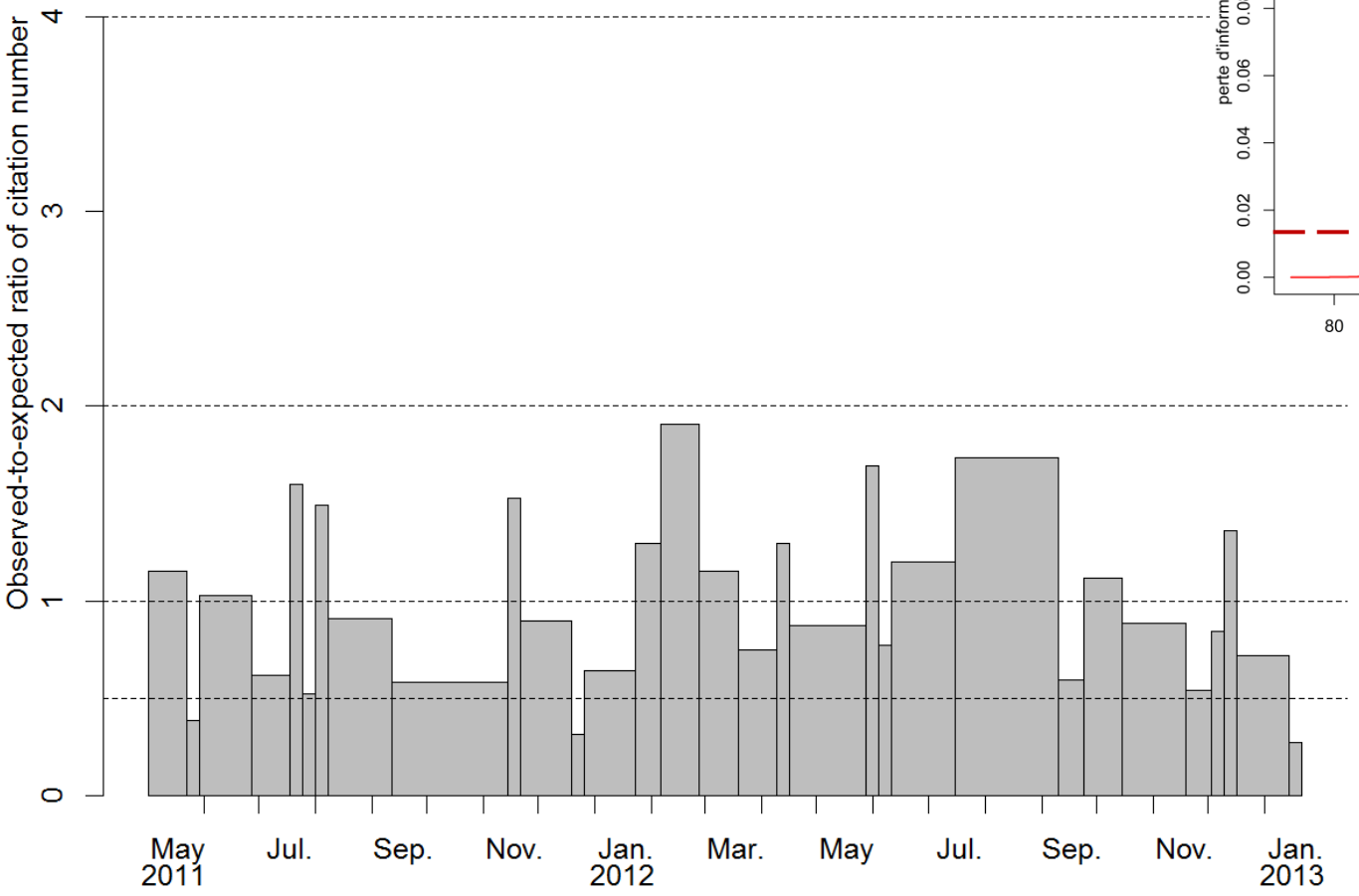
[Giraud, Grasland, Lamarche-Perrin *et al.*, ECTQG 2013]

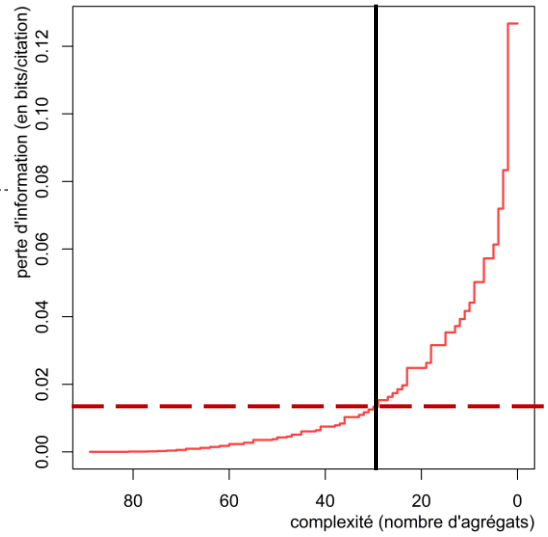




La Syrie vue par LE MONDE

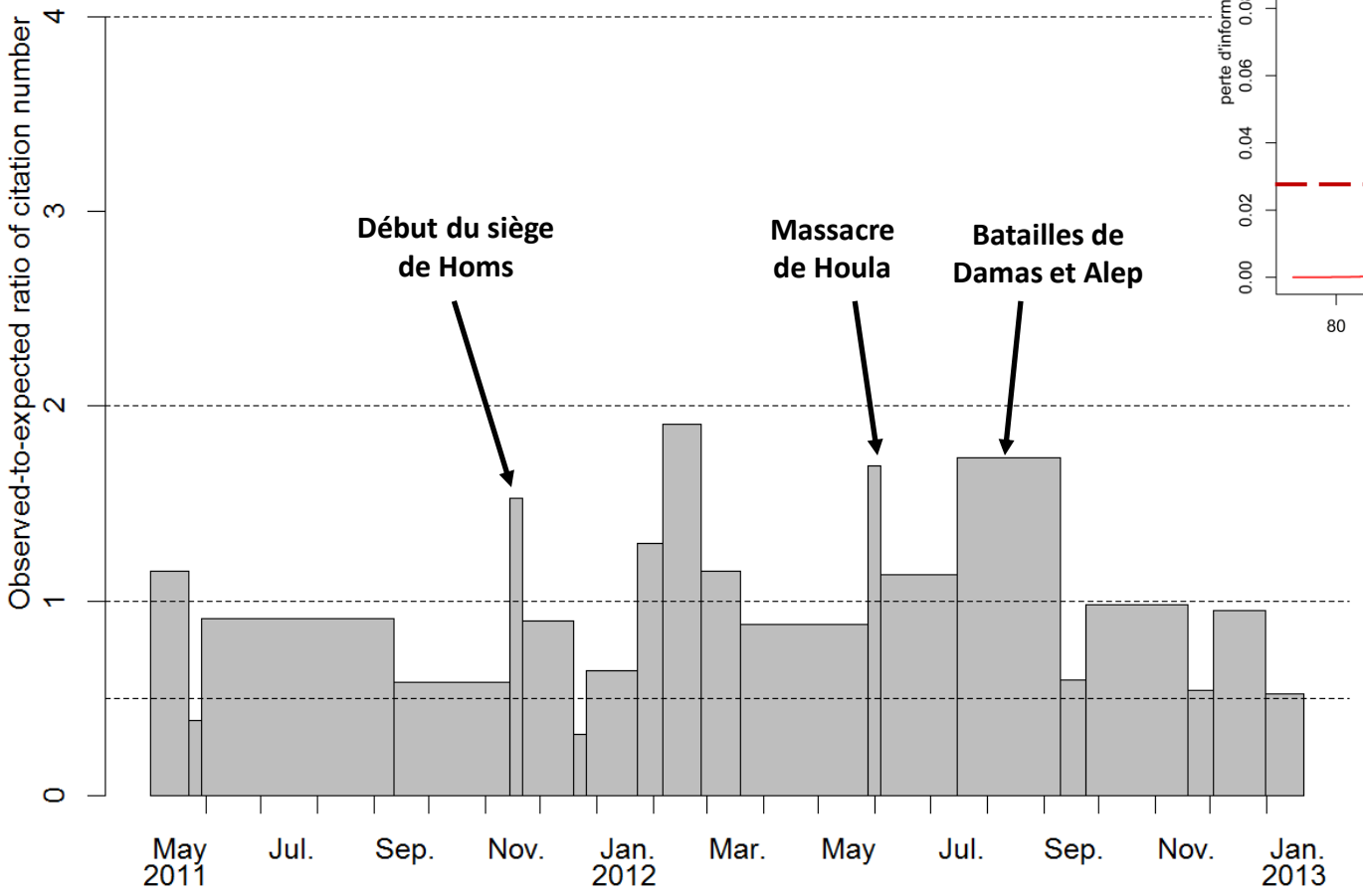
[Giraud, Grasland, Lamarche-Perrin *et al.*, ECTQG 2013]

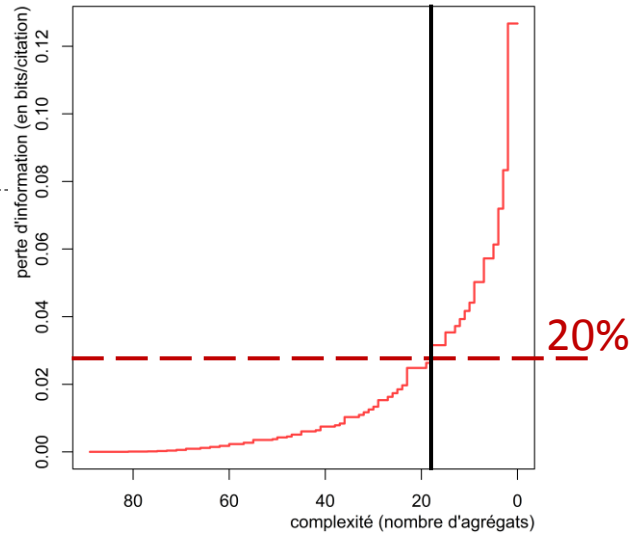




La Syrie vue par LE MONDE

[Giraud, Grasland, Lamarche-Perrin *et al.*, ECTQG 2013]

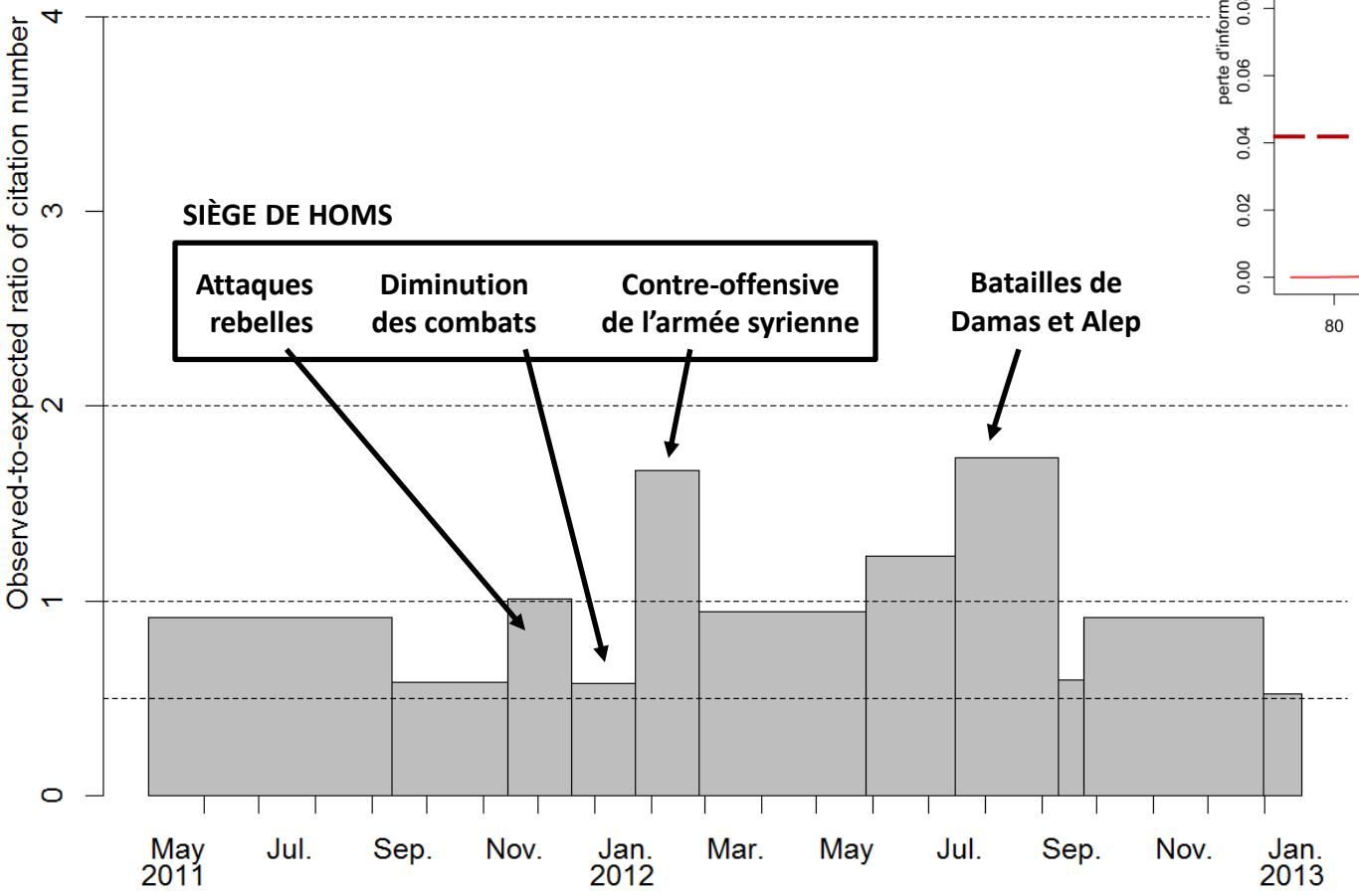


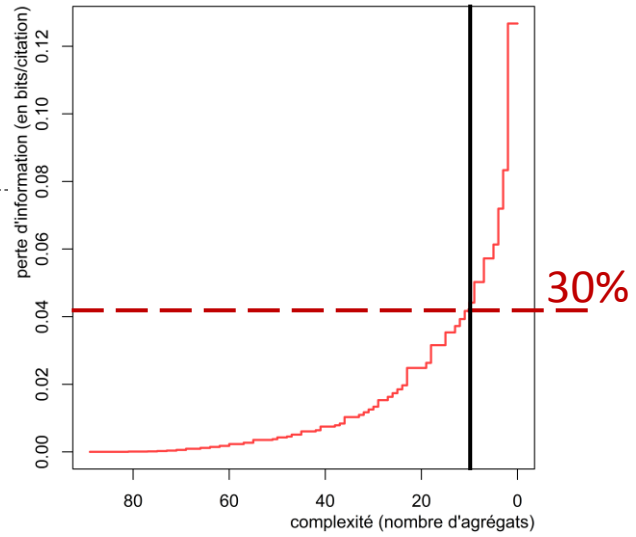


La Syrie vue par LE MONDE

[Giraud, Grasland, Lamarche-Perrin *et al.*, ECTQG 2013]

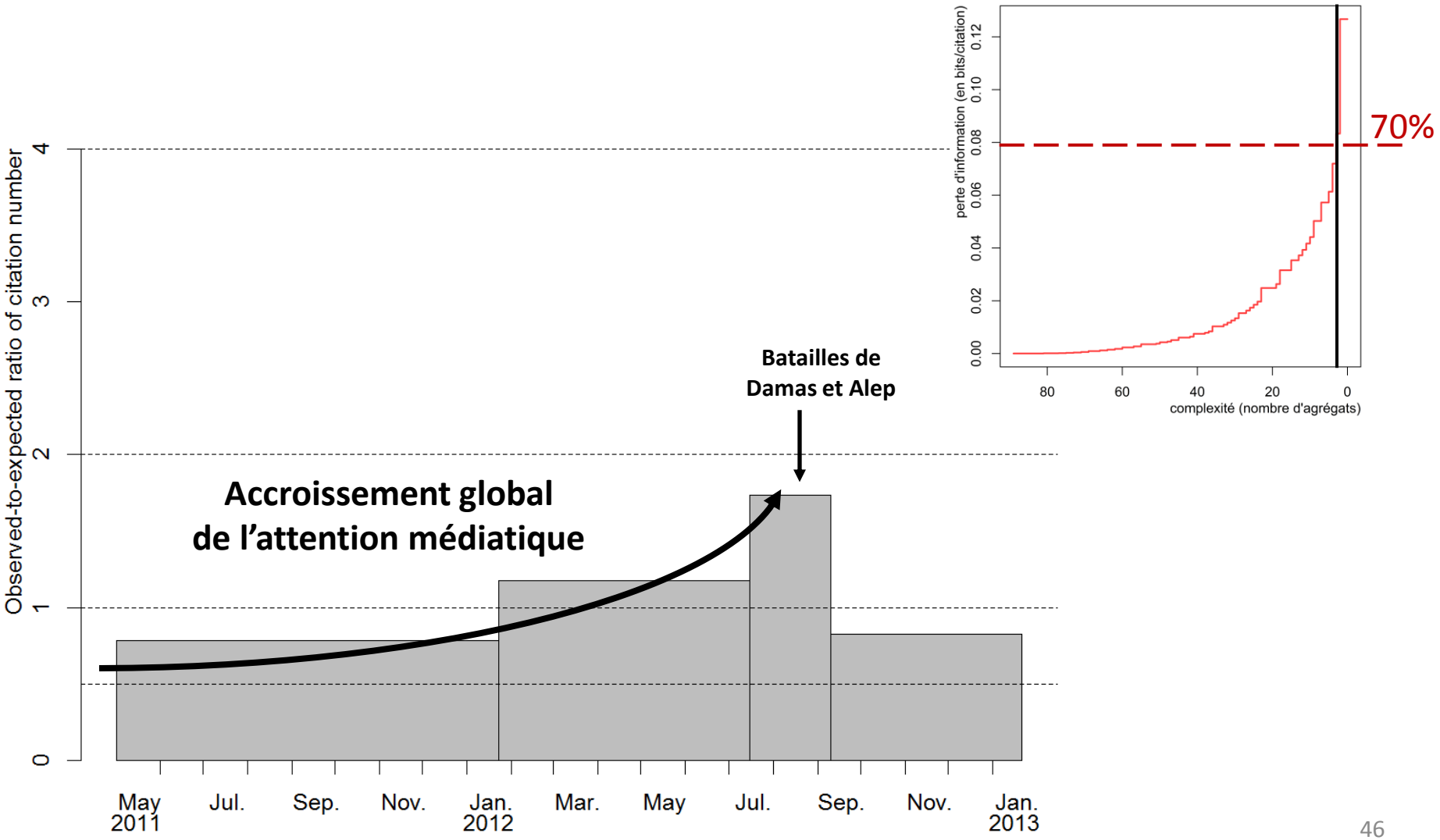
Source : Wikipedia
Timeline of the Syrian civil war
Siege of Homs





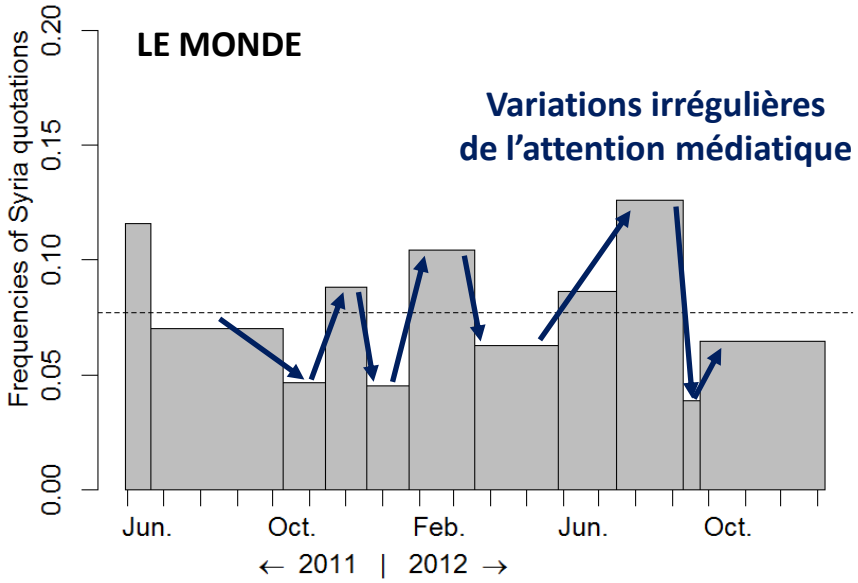
La Syrie vue par LE MONDE

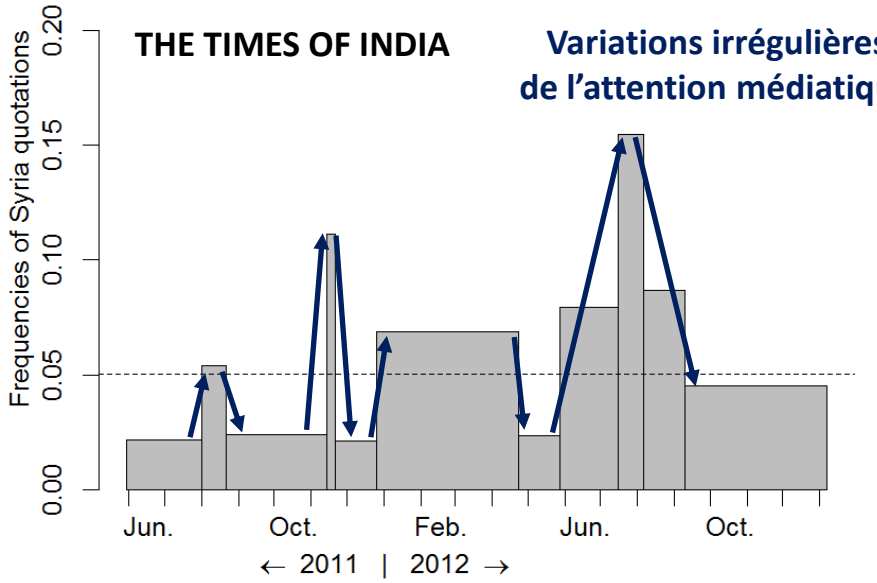
[Giraud, Grasland, Lamarche-Perrin *et al.*, ECTQG 2013]

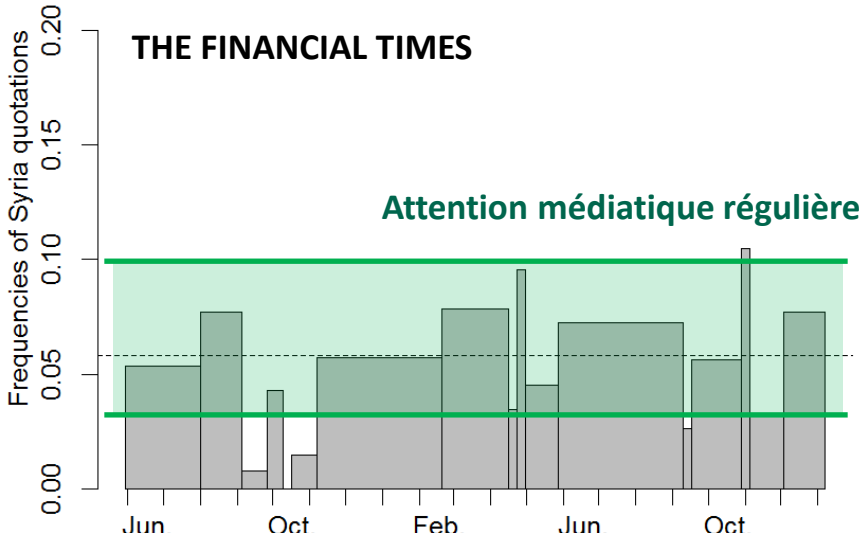


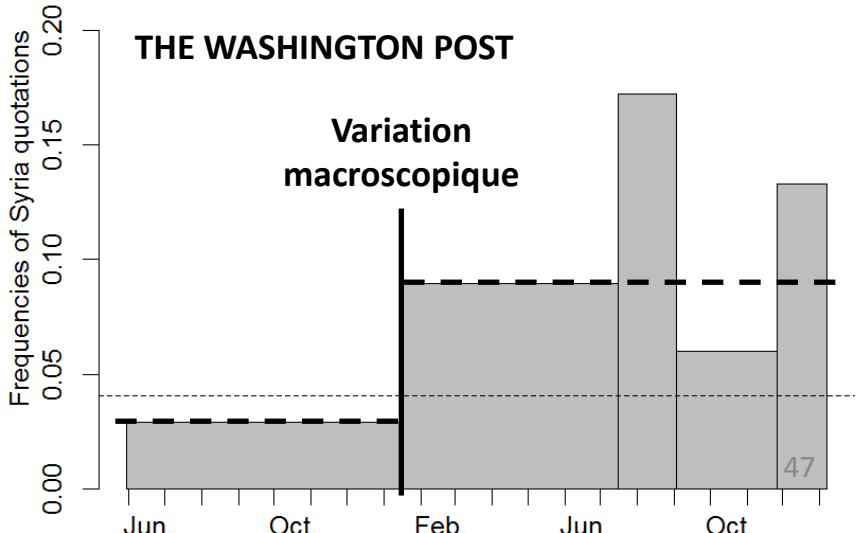
La Syrie vue par 4 journaux

[Giraud, Grasland, Lamarche-Perrin *et al.*, ECTQG 2013]

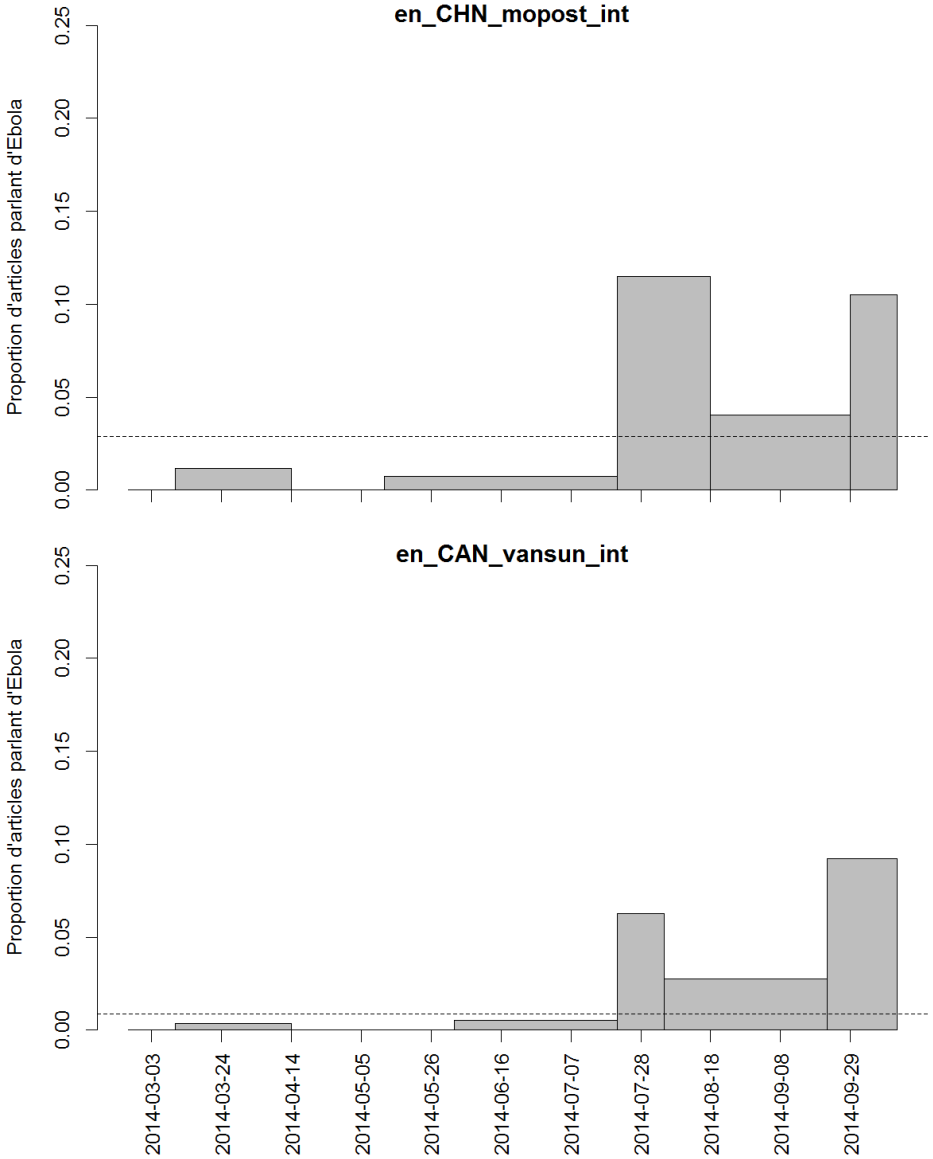


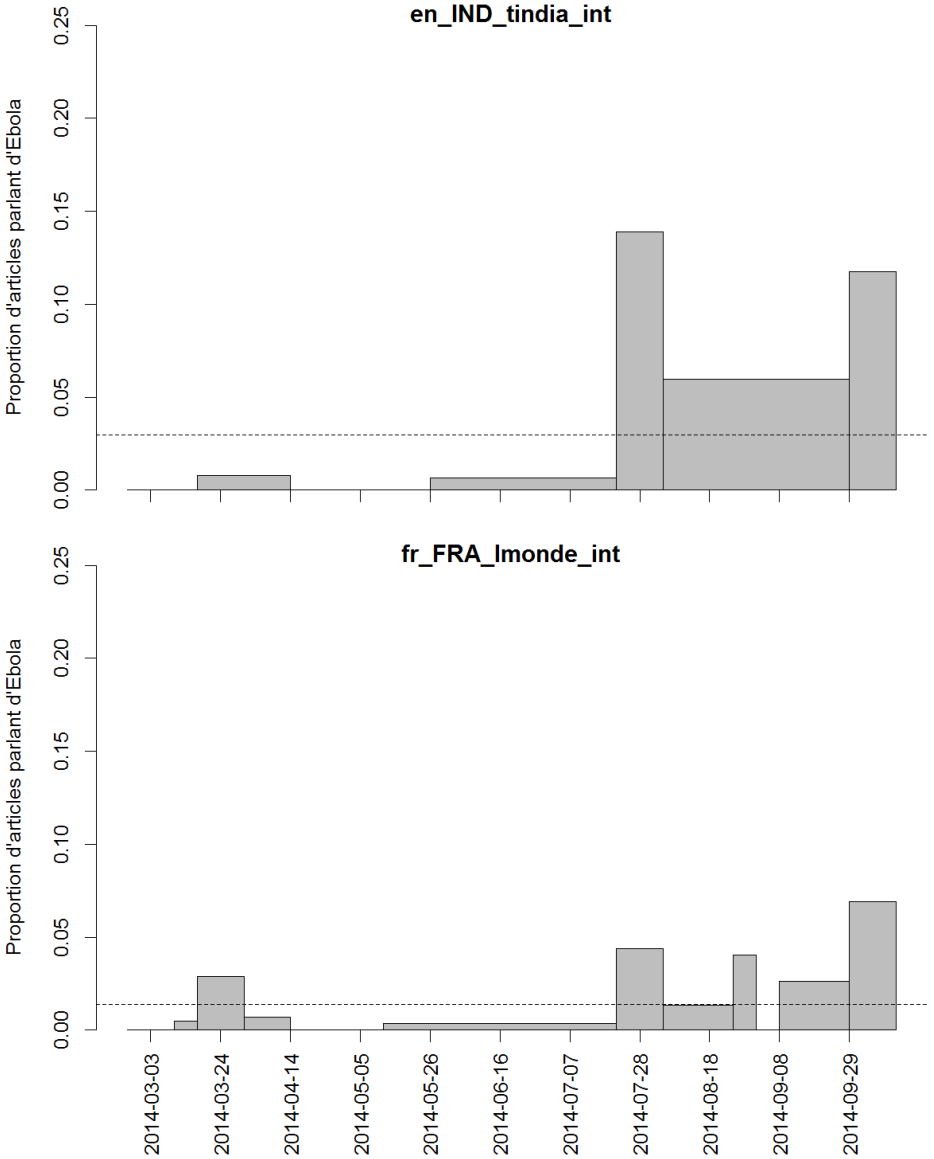




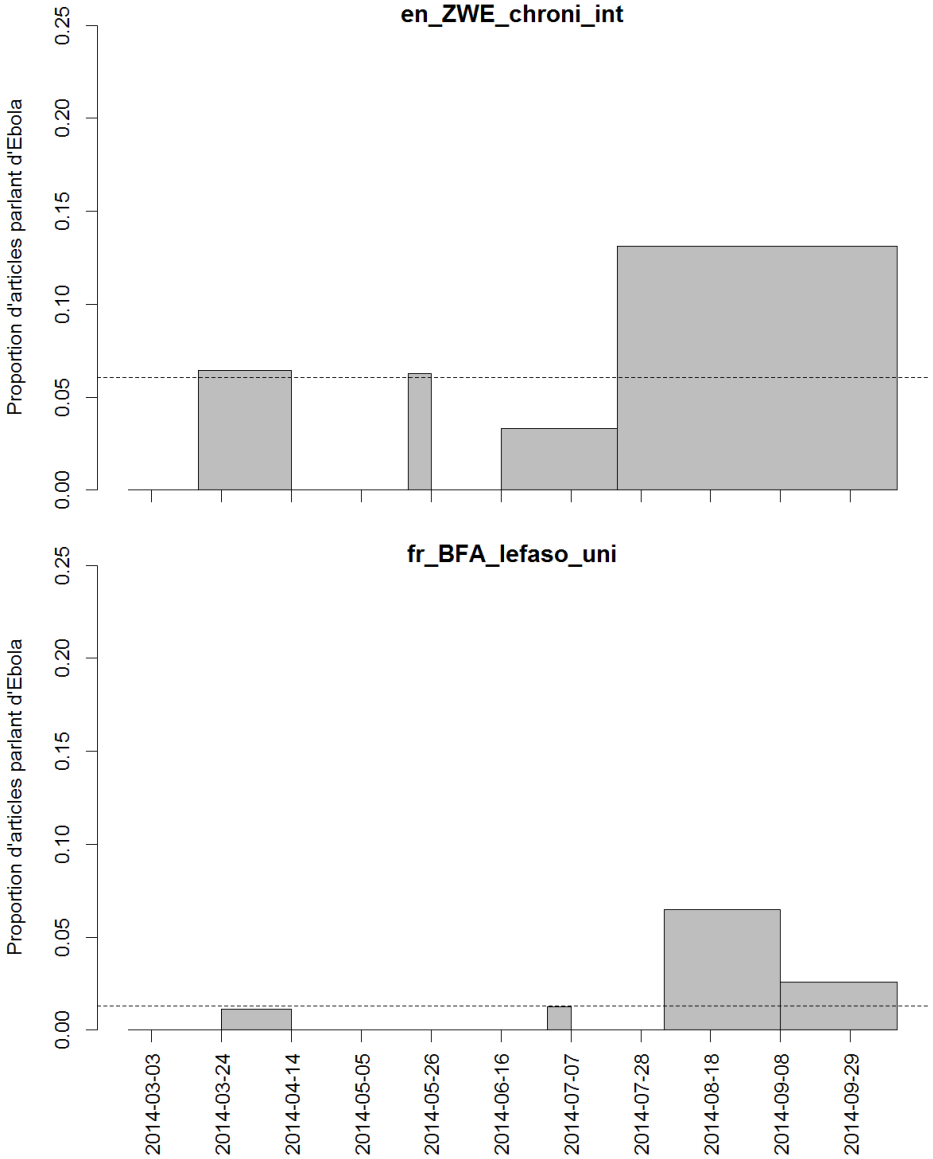


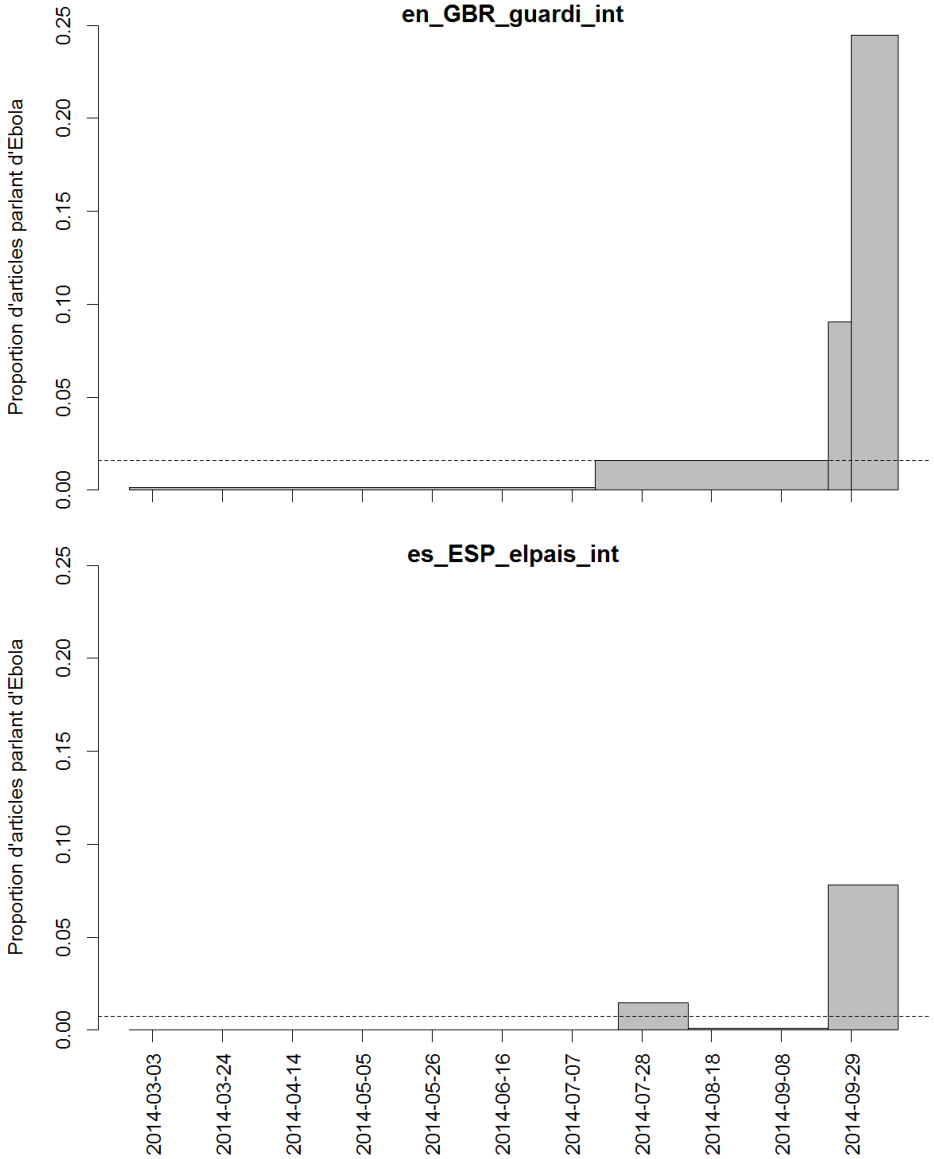
Ebola vu par 8 journaux (90% d'information)



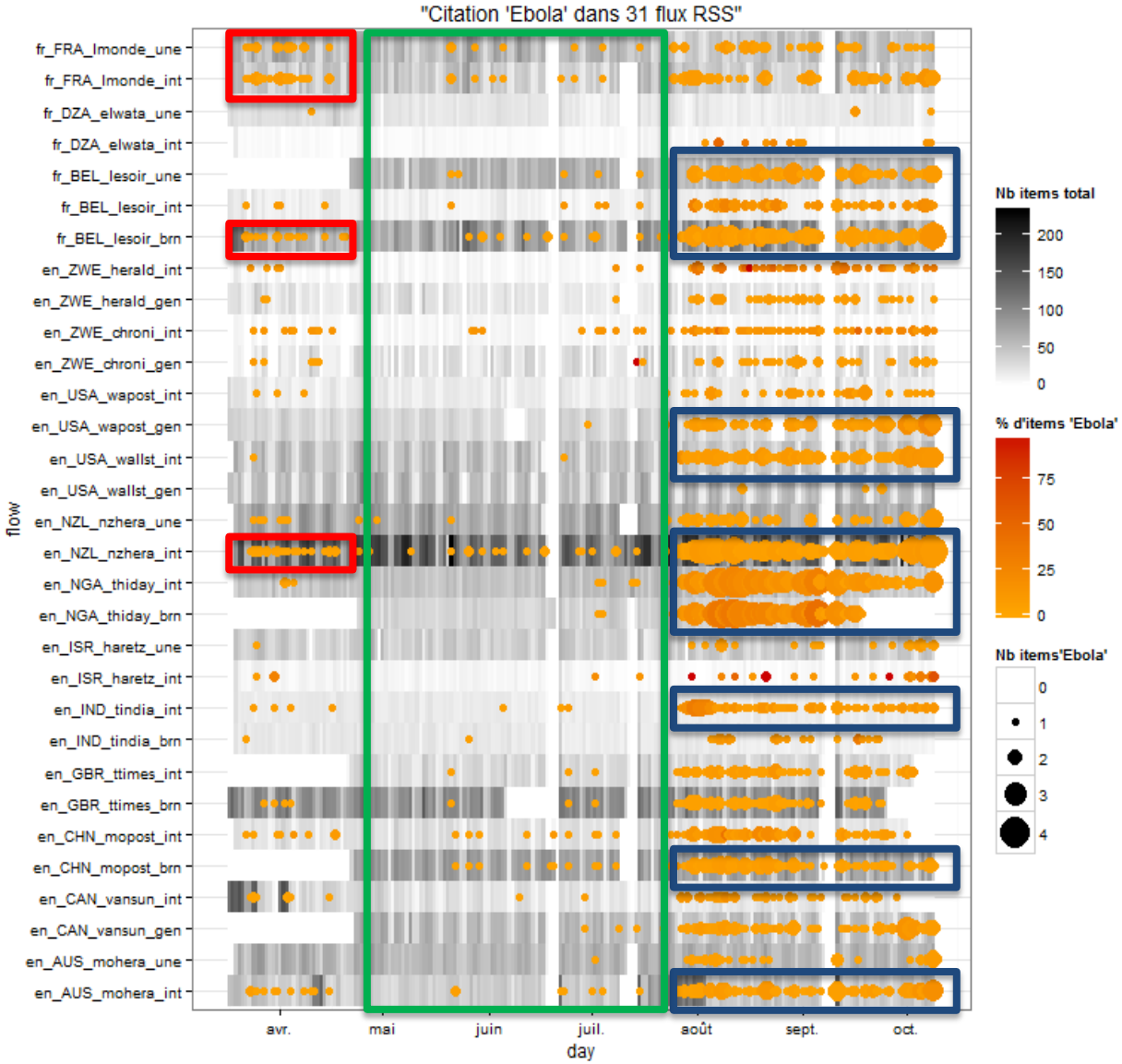


Ebola vu par 8 journaux (90% d'information)

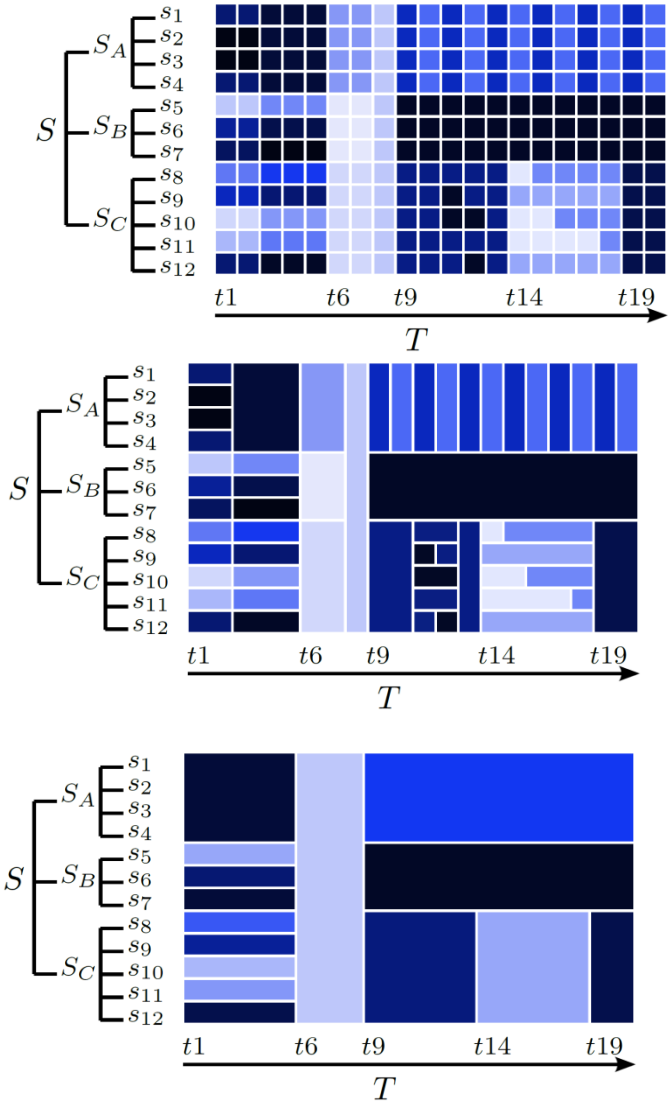


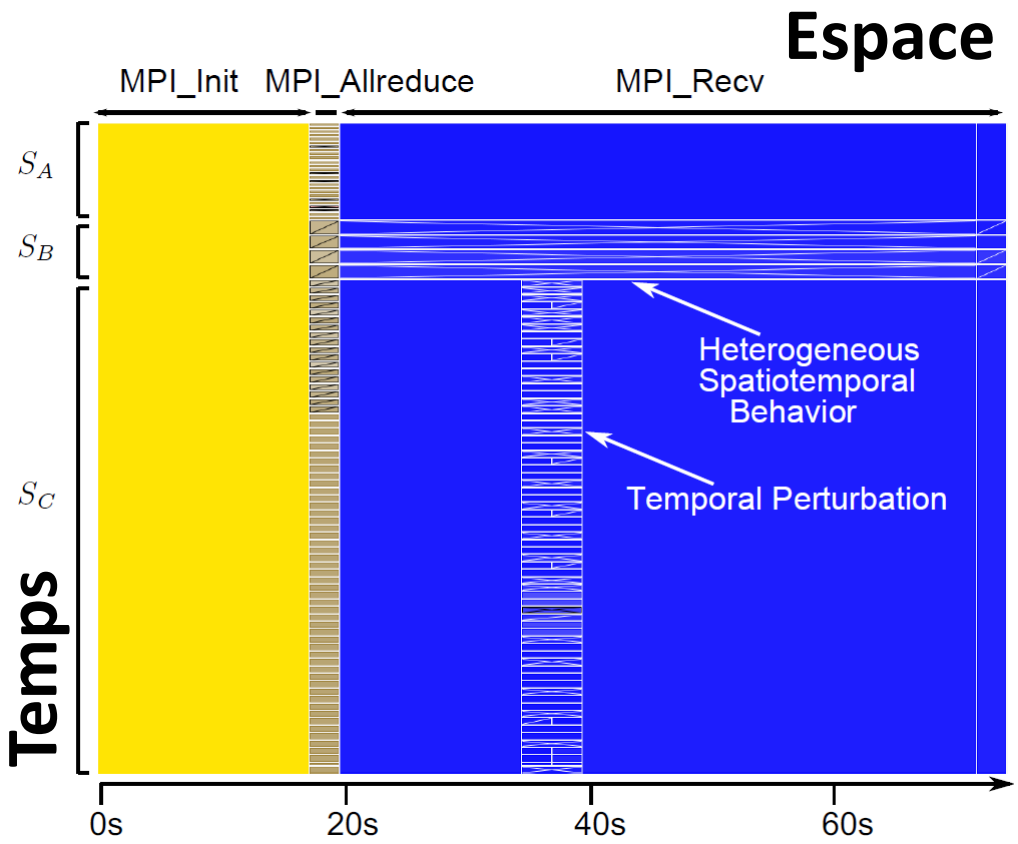


Agrégation médiatique

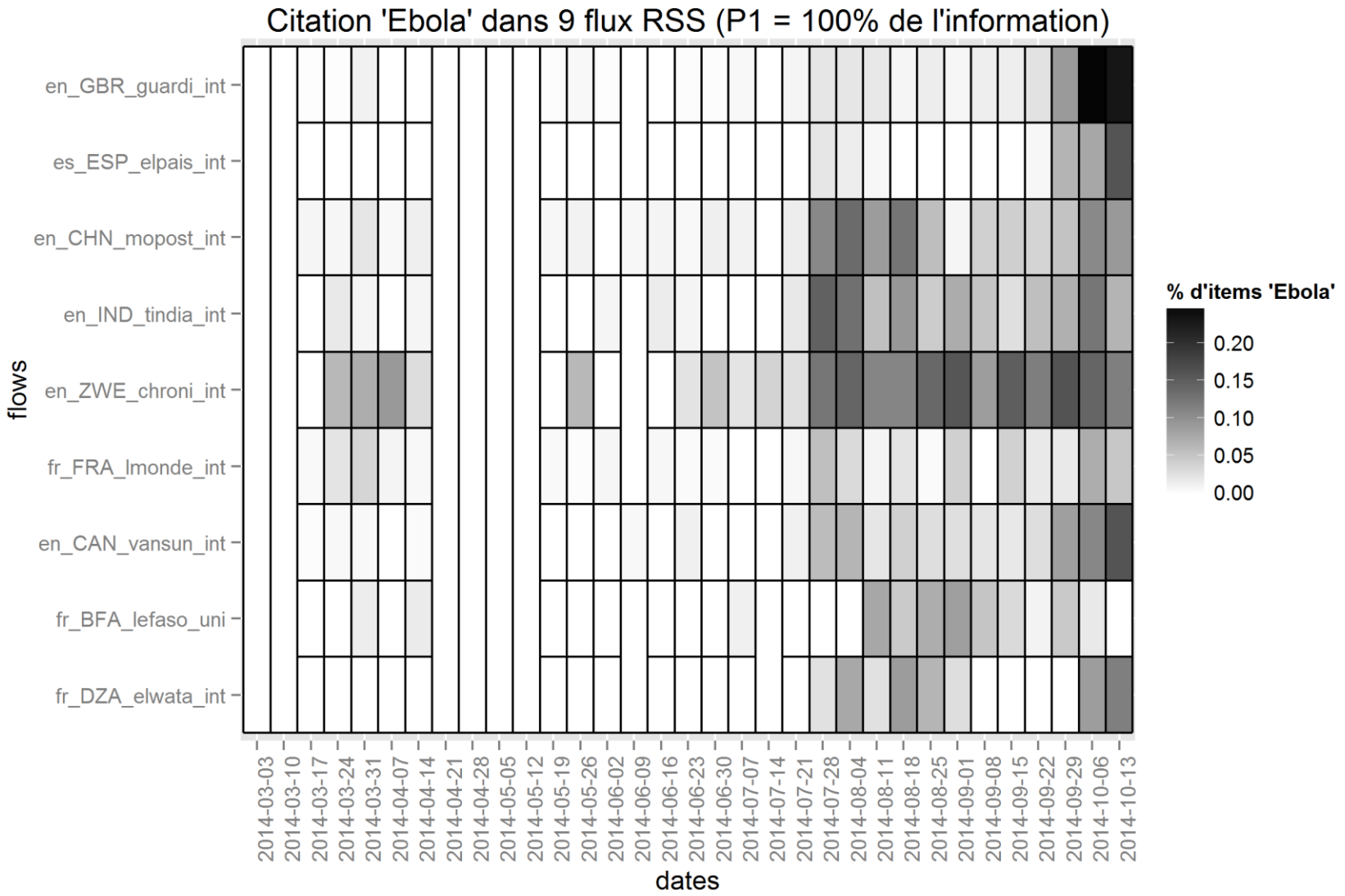


Agrégation spatiotemporelle

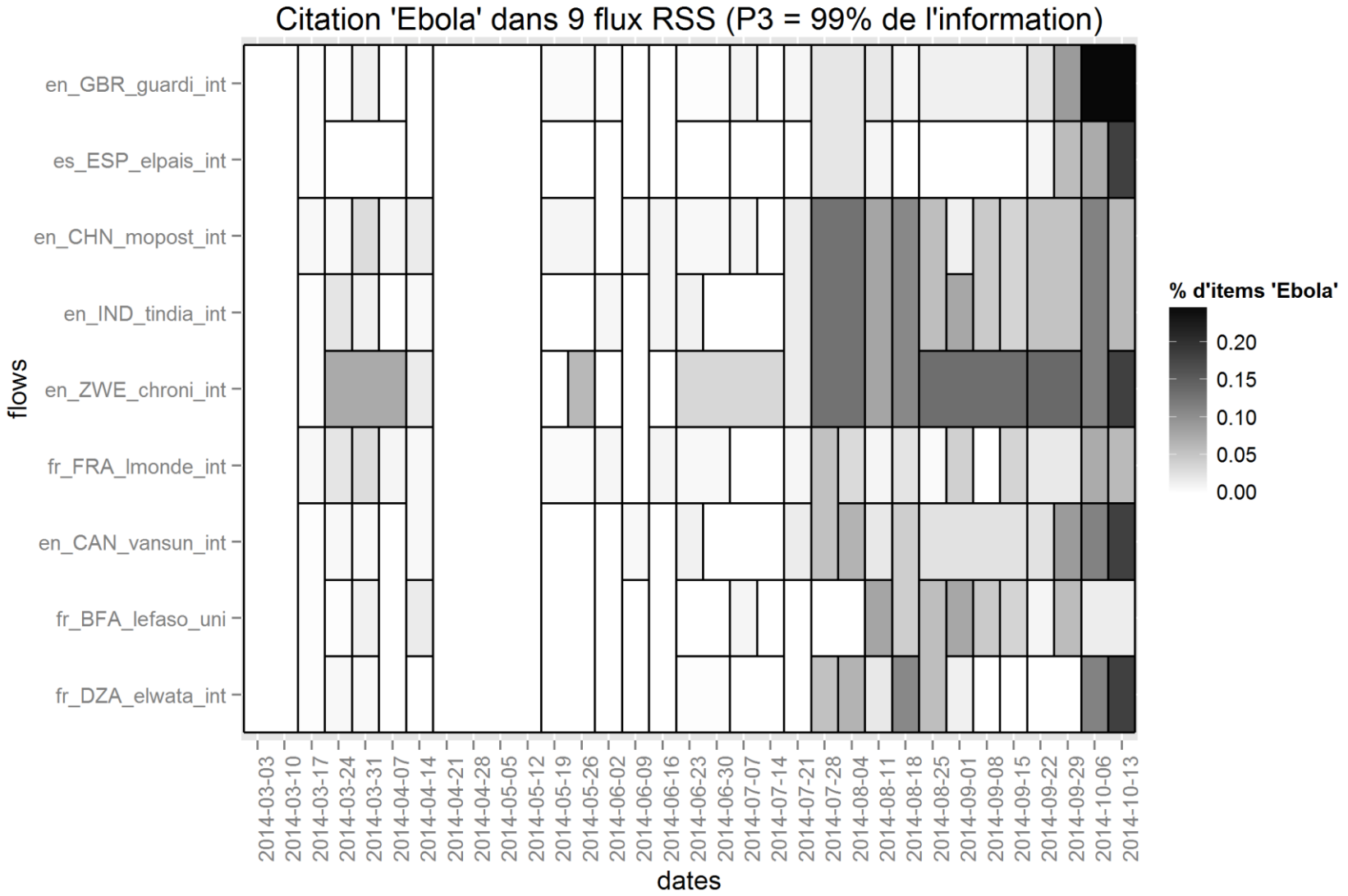




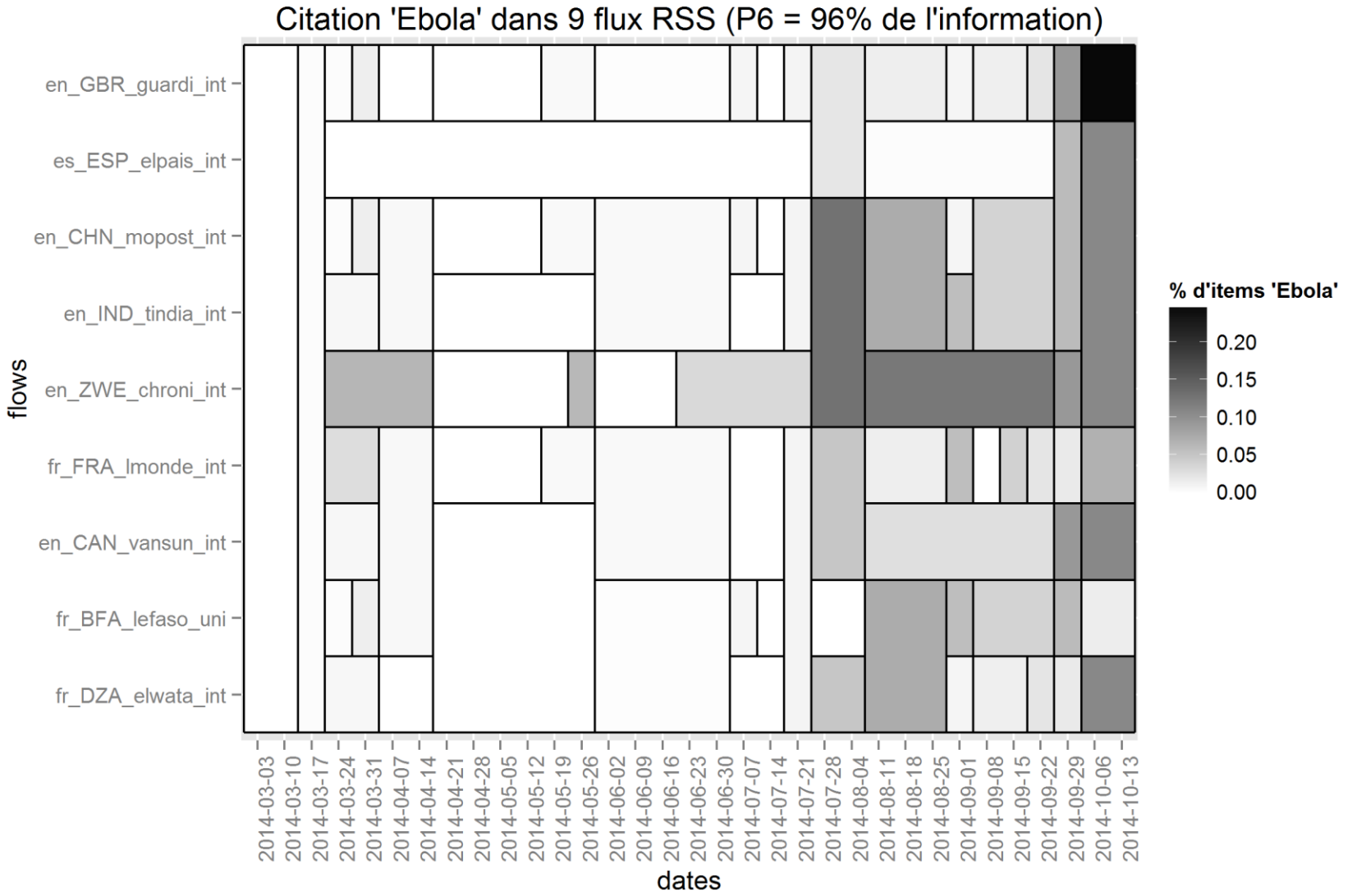
Agrégation médiatique



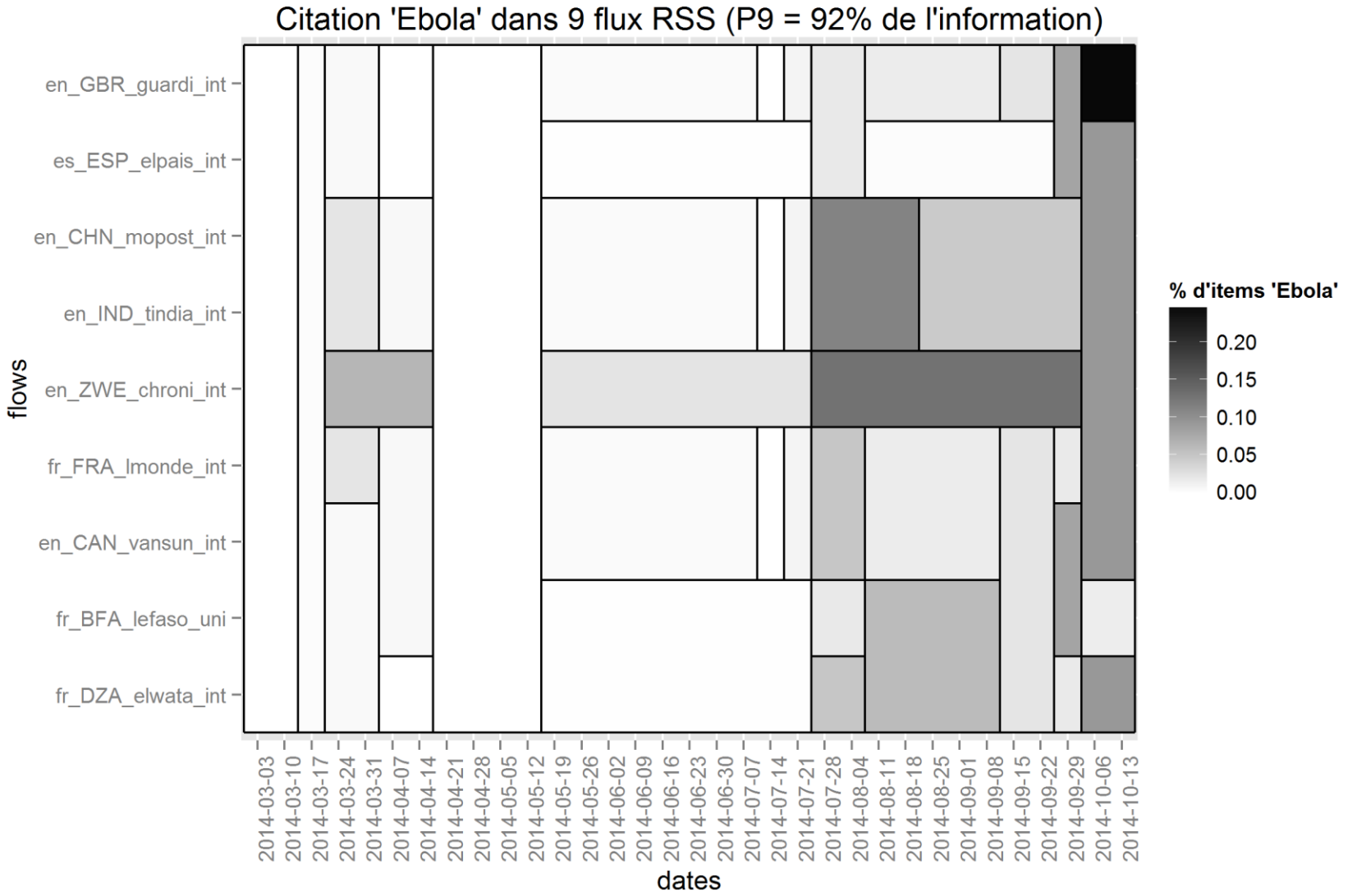
Agrégation médiatique



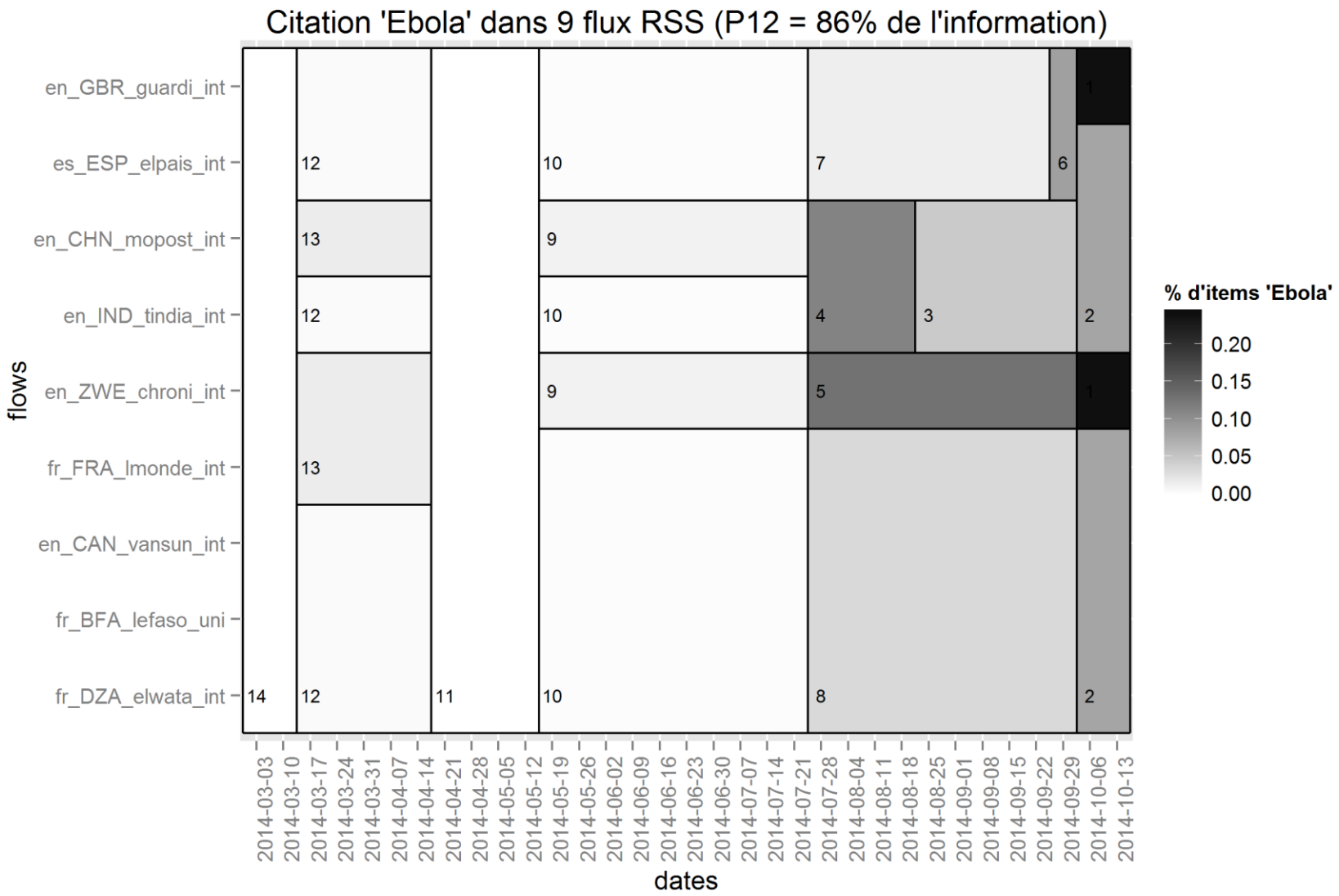
Agrégation médiatique



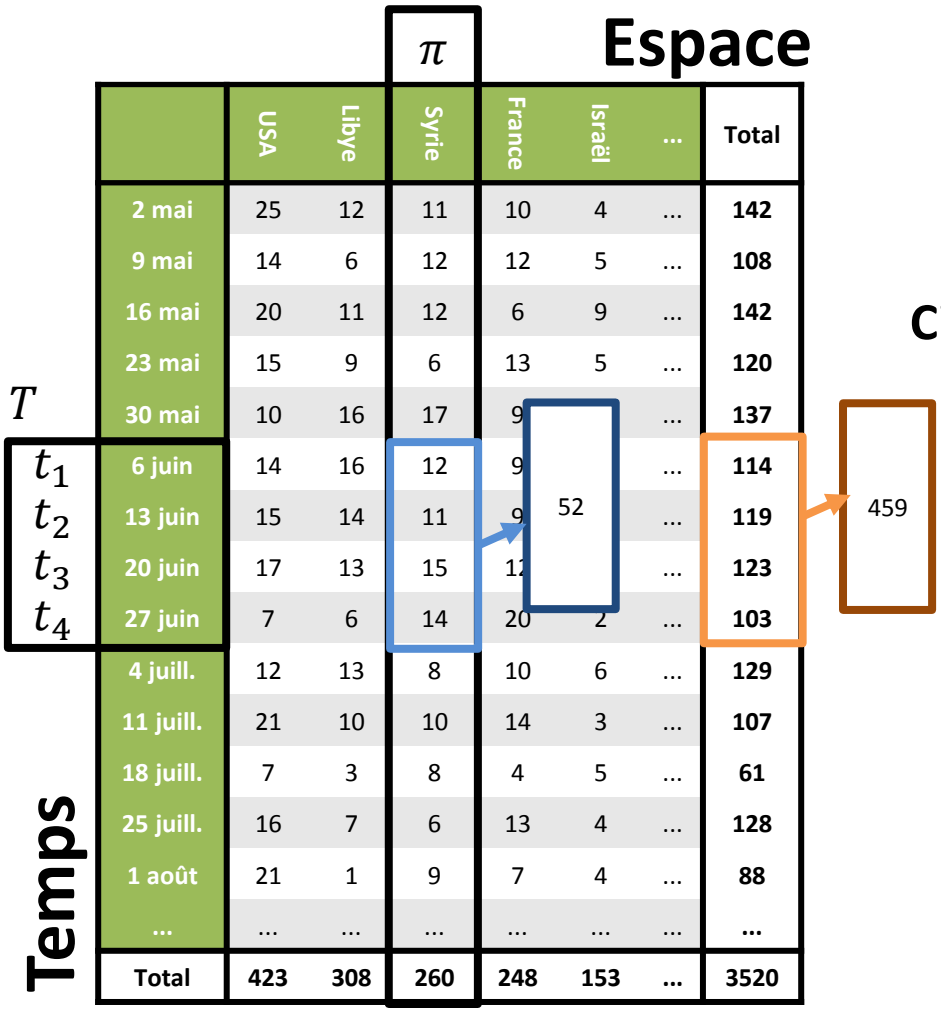
Agrégation médiatique



Agrégation médiatique



Agrégation et modèle de données



Citations réelles

$v(\pi, t_1) = 12$
$v(\pi, t_2) = 11$
$v(\pi, t_3) = 15$
$v(\pi, t_4) = 14$

Valeurs marginales

$v(., t_1) = 12$
$v(., t_2) = 11$
$v(., t_3) = 15$
$v(., t_4) = 14$

$v(., T) = 459$

Citations observées

$v(\pi, T) = 52$

Modèle 1

$$v^*(\pi, t) = \frac{v(\pi, T)}{|T|}$$

$v^*(\pi, t_1) = 13$
$v^*(\pi, t_2) = 13$
$v^*(\pi, t_3) = 13$
$v^*(\pi, t_4) = 13$

Modèle 2

$$v^*(\pi, t) = v(\pi, T)\frac{v(., t)}{v(., T)}$$

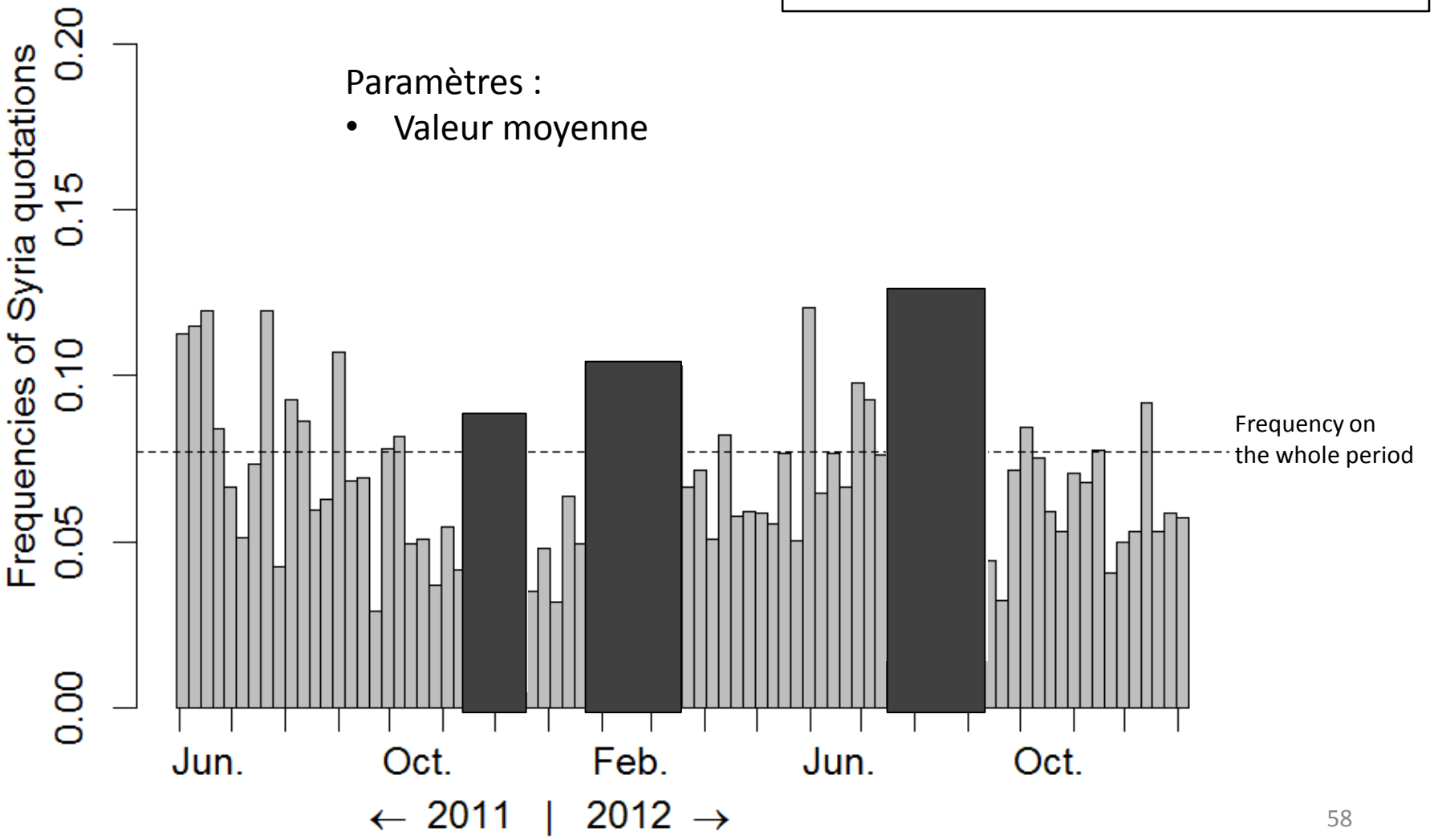
$v^*(\pi, t_1) = 11.8$
$v^*(\pi, t_2) = 11.2$
$v^*(\pi, t_3) = 15.1$
$v^*(\pi, t_4) = 13.9$

Agrégation et modèle de données

Journal LE MONDE concernant la Syrie

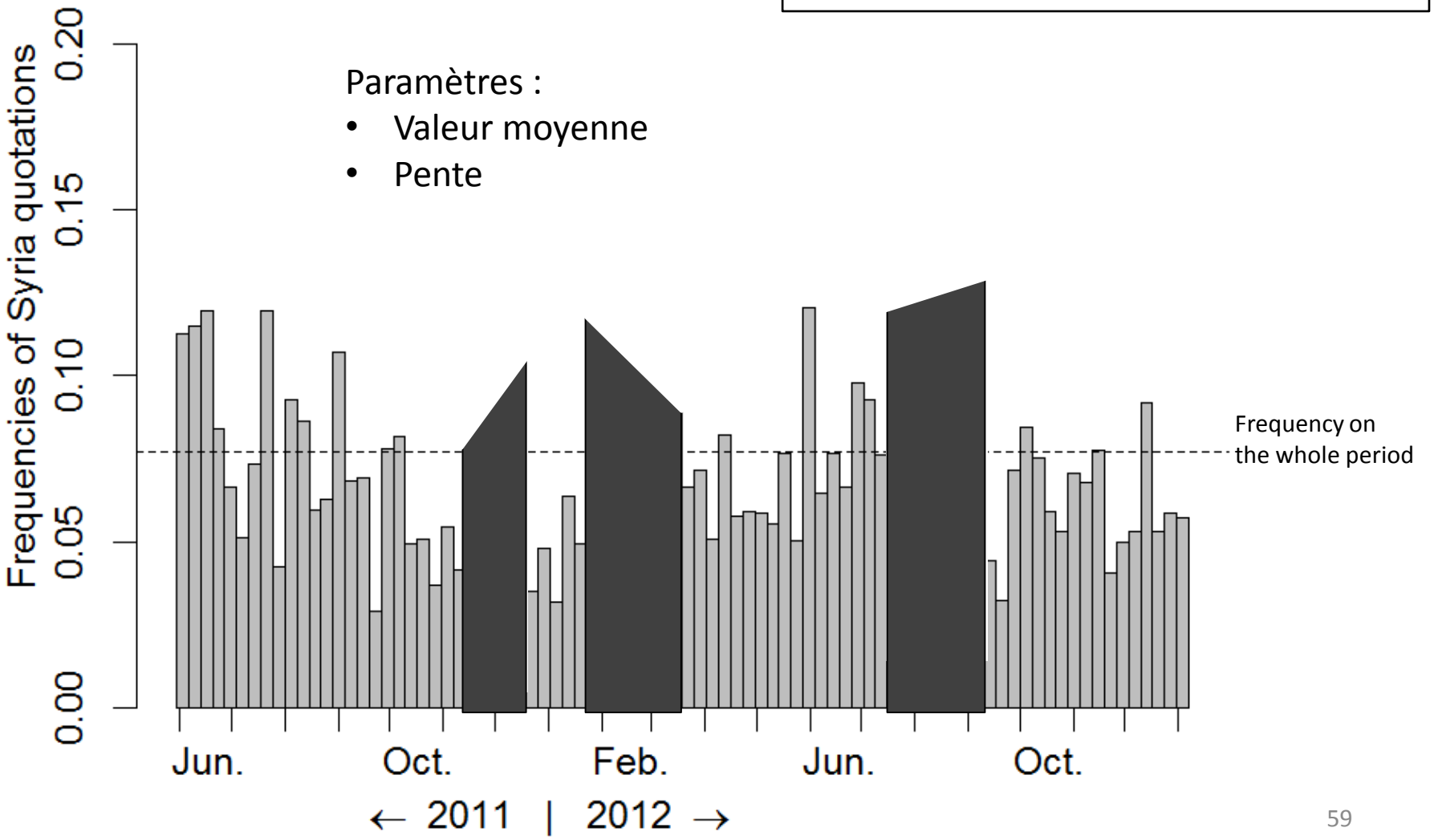
Paramètres :

- Valeur moyenne



Agrégation et modèle de données

Journal LE MONDE concernant la Syrie



ANR CORPUS GEOMEDIA

Paris, 25 novembre 2014

MERCI POUR VOTRE ATTENTION

Courriel: Robin.Lamarche-Perrin@mis.mpg.de
Web: www.mis.mpg.de/jjost/members/robin-lamarche-perrin.html