Impact of scales and sample size on temperature assessment within rescaling processes

Pierre-Alain Ayral*, <u>Didier Josselin</u>*, Céline Lacaux, Nicolas Martin*, Matthieu Vignal***

Observatoire Hydro-météorologique Méditerranéen Cévennes Vivarais

ZABR

LMA Laboratoire de Mathématiques d'Avignon

* UMR ESPACE, CNRS, France
** Laboratoire de Mathématique d'Avignon, France

ESPACE

Outline

- The change of (spatial) support problem
- Upscaling temperatures using different spatial partitions in the PACA region
- Temperature downscaling in PACA
- Conclusion

The change of support problem

- Modifiable Areal Unit Problem
- Simpson paradox
- Robinson paradox
- Ecological fallacy

	Sick after pesticides	OK after pesticides	
Apollon Butterfly	200	800	*1000*
Bonali Eagle	50	950	*1000*
	250	*1750*	*2000*

Probability to be sick for Butterfly : 200/1000 = 0.20

Probability to be sick for Eagle : 50/1000 = 0.05

Relative risk = 0.20/0.05 = 4

(4 times more risk for Butterfly)

Total	sick	OK	
Butterfly	200	800	*1000*
Eagle	50	950	*1000*
	250	*1750*	*2000*

RR = (200/1000) / (50/1000) = 4.0

This is the Simpson paradoxe

Group 1	sick	OK	
Butterfly	193	224	*417*
Eagle	39	45	*84*
	232	*269*	*501*

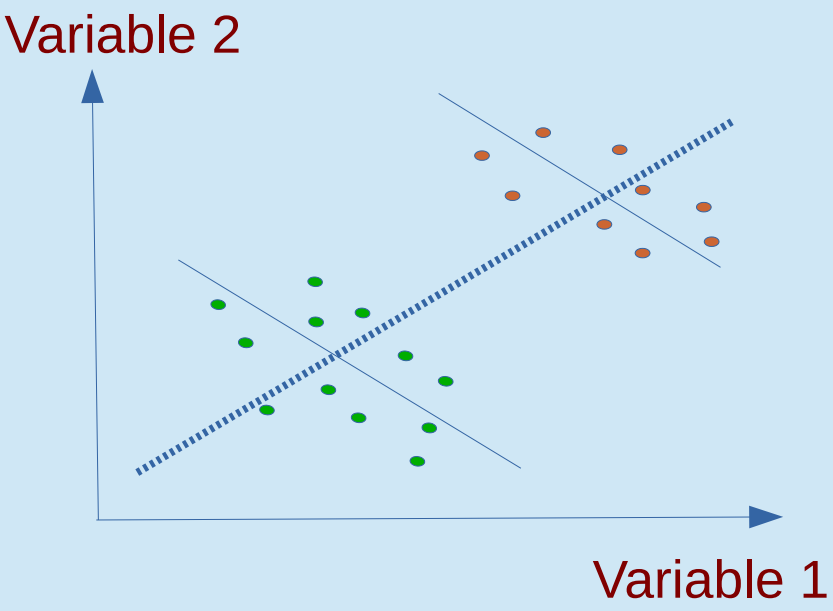
RR = (193/417) / (39/84) = 1.0

Group 2	sick	OK	
Butterfly	7	576	*583*
Eagle	11	905	*916*
	18	*1481*	*1499*

RR = (7/583) / (11/916) = 1.0

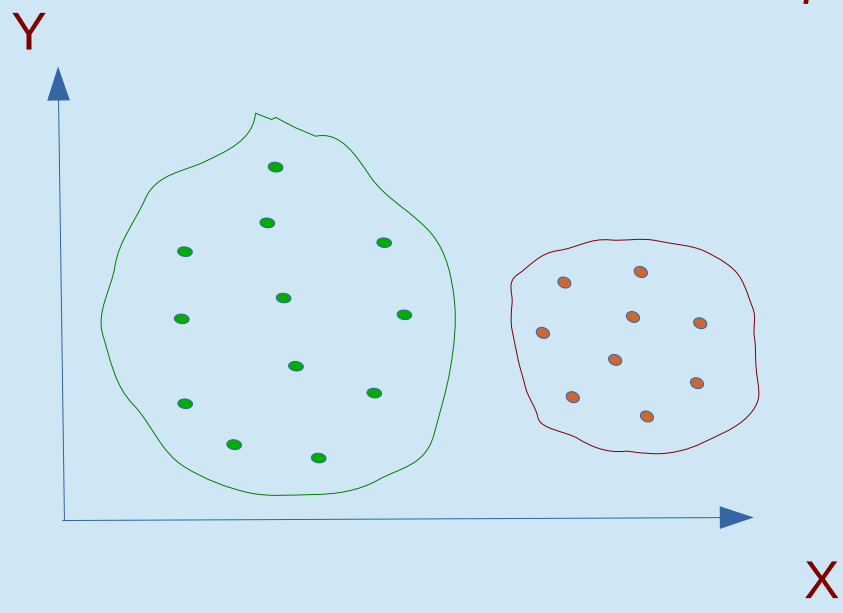
A problem of (spatial) clustering

2 clusters of data



In a plot

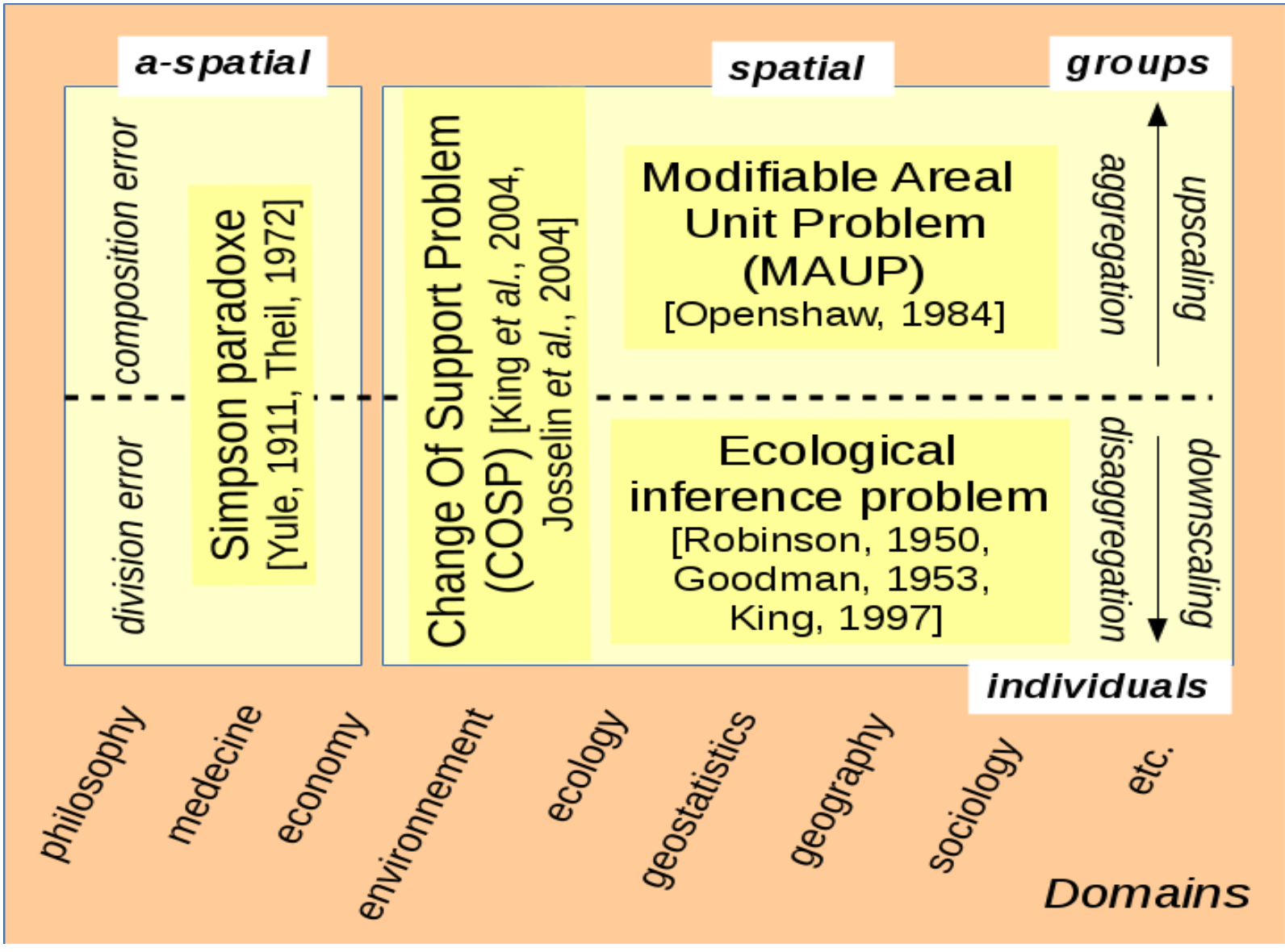
2 clusters of points



In space

Scales effects, upscaling, downscaling

(Josselin & Louvet 2016)



Rationalité

Postulats

- Un indicateur statistique mesuré sur des agrégats de valeurs est hautement dépendant de la structure de ces agrégats
- Il est possible d'extraire l'effet de cette structure des agrégats par des méthodes de redistribution aléatoire de type permutations
- En géographie, cela se traduit par l'erreur écologique (*ecological fallacy*), l'effet du support spatial (*COSP*) et le problème des entités surfaciques modifiables (*MAUP*) sur tout indicateur
- Le processus observé dépend beaucoup des effets des variances *inter* et *intra* des agrégats liés au découpage (effets inverses de taille)

Définitions

- **Entité atomique** : plus petit grain spatial (parcelle cadastrale)
- **Agrégat** : regroupement d'entités atomiques (cantons, EPCI, département...)
- **Partition spatiale** : découpage administratif complet et continu composé d'agrégats
- **Indicateur** : estimateur statistique appliqué sur les valeurs des entités atomiques dans chaque agrégat :
  - ***Valeurs centrales*** : moyenne, médiane
  - ***Indices de dispersion*** : écarts-types, MAD
- **Agrégateur** : estimateur appliqué sur les valeurs des indicateurs des agrégats d'une partition
- **Facteur** : dimension ou composante de la valeur mesurée (à extraire, dont particulièrement l'effet du support spatial)
- **Méthode** : méthode de traitement des données observées :
  - ***Sans ré-échantillonnage*** (« Global », « Obs »)
  - ***Avec ré-échantillonnage*** (« Alea », « Alea-tot », « Alea-sel »)

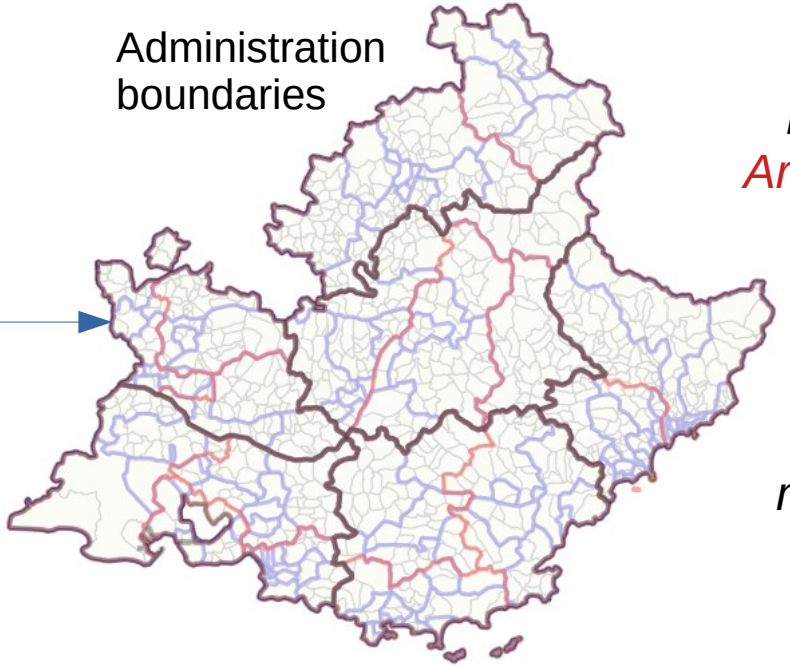
Aggregating temperatures using different spatial partitions in the PACA region

- Administration boundaries and partitioning
- Meteorological stations
- ALADIN model grid
- Temperature measurement

Combining geographical data

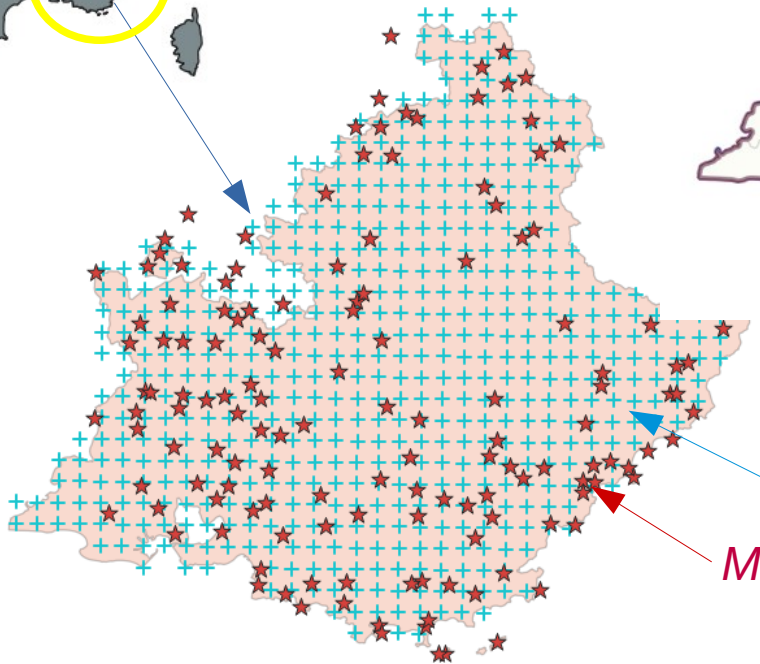
PACA region
(continental part)

Administration boundaries



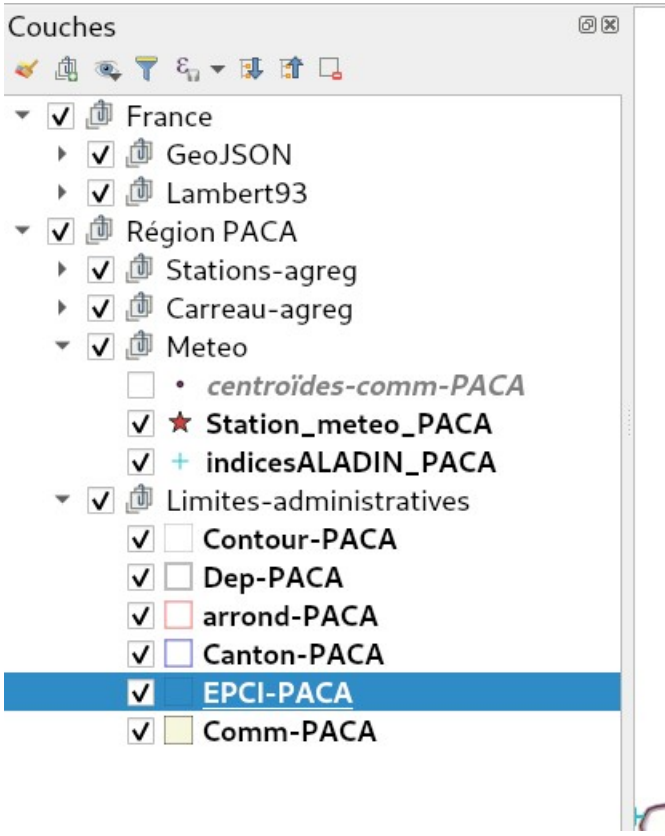
(French)
Département
Arrondissement
Canton
Commune
+
EPCI
(groups of municipalities)

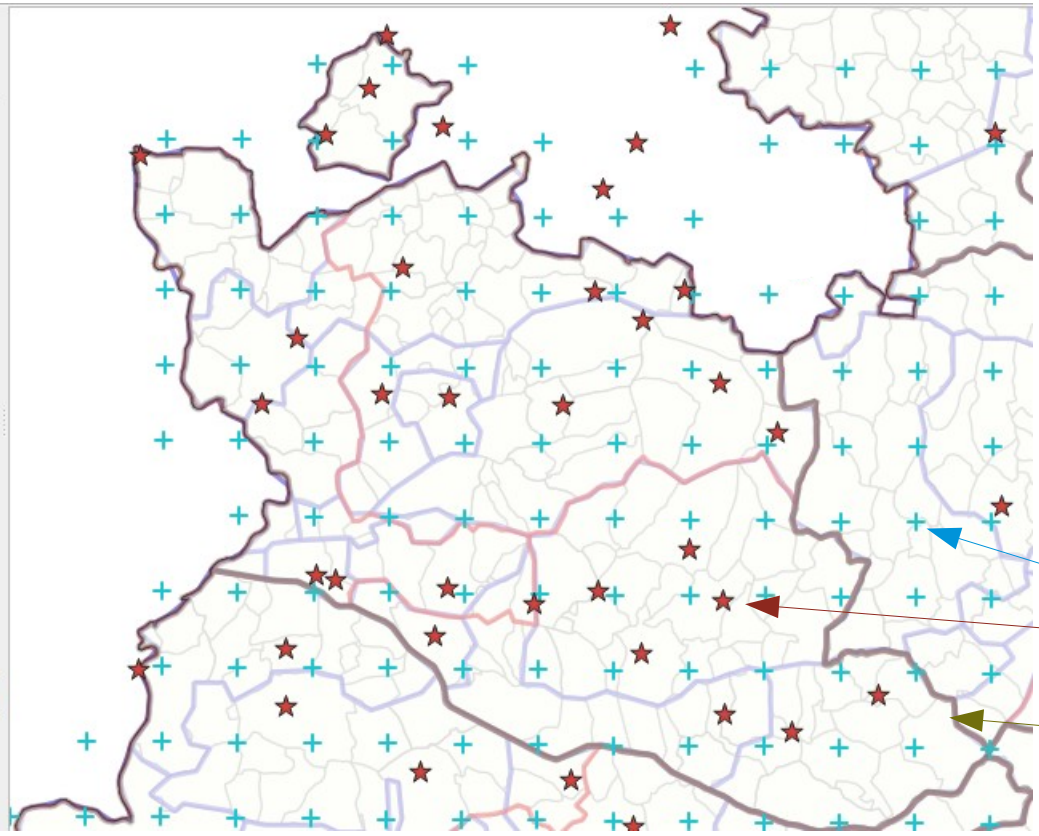
100 km



ALADIN model grid (8 x 8 km)
Meteorological stations (Meteo France)

Statistical (up)scaling process





For **each spatial entity** at **different Imbricated scales**, we process a statistical **aggregation** (*mean, median, standard deviation…*) on **averaged temperatures** provided on :

- *the grid points*
- *the local stations*

e.g. a spatial entity: the Département of Vaucluse

Distribution observée avec partition (avec agrégation)

On calcule l'indicateur sur la distribution de données observées, sur chaque partition

9 cellules
en 3 agrégats

Température

Agrégats
1, 2 et 3

Moyenne des T° de l'agrégat 3
= 26 °C

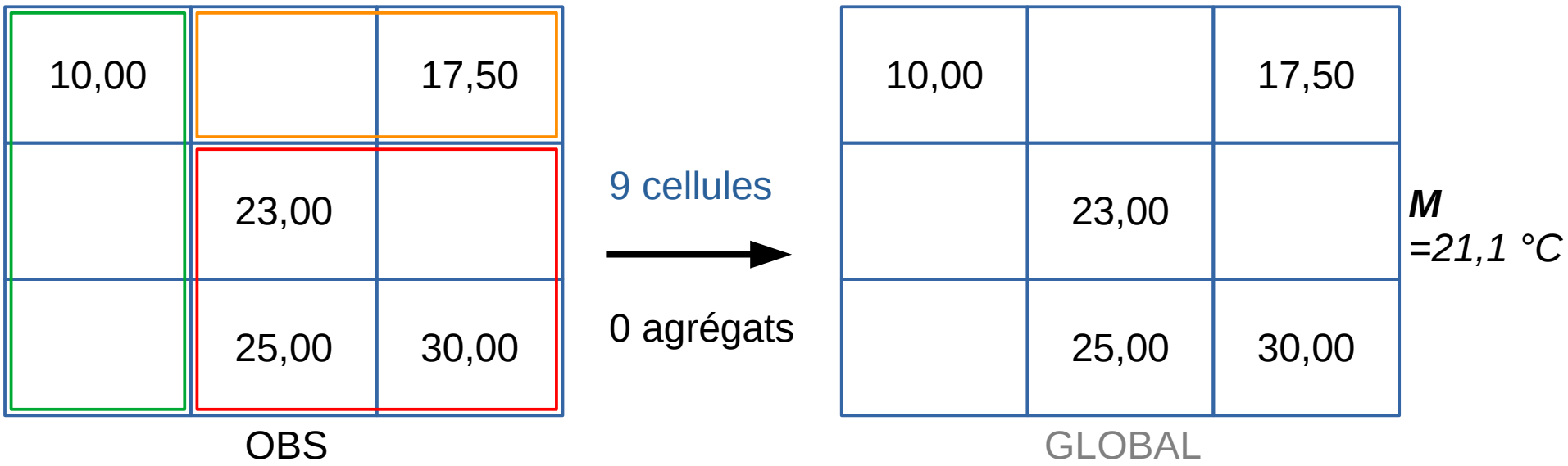
OBS

N=5
N_Agreg=3

Exemple :
Moyenne des moyennes M
= (10,00+17,50+26,00)/3
= 17,83 °C

Distribution observée sans partition (mêmes distribution statistique et périmètre)

Pas de ré-échantillonnage, on calcule l'indicateur sur la distribution de données observées sans partition



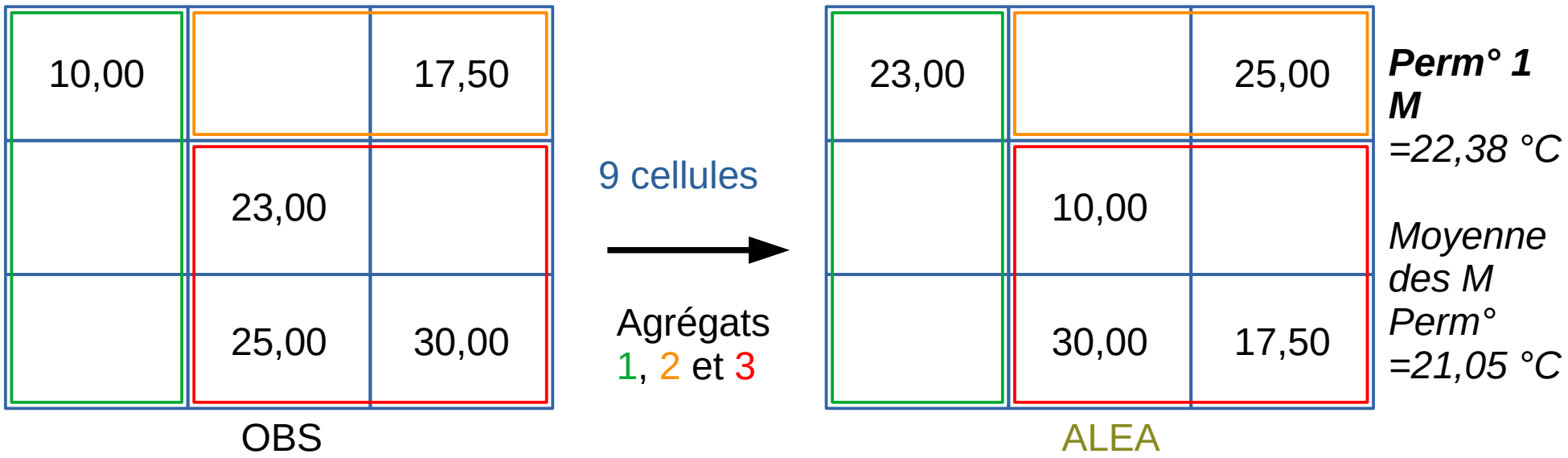
N=5
N_Agreg=3

Distribution spatiale des valeurs différentes

N=5
N_Agreg=0

Distribution aléatoire à topologie et distribution statistique équivalentes

Les mêmes parcelles sont affectées de mutations observées, tirées au hasard sans remise



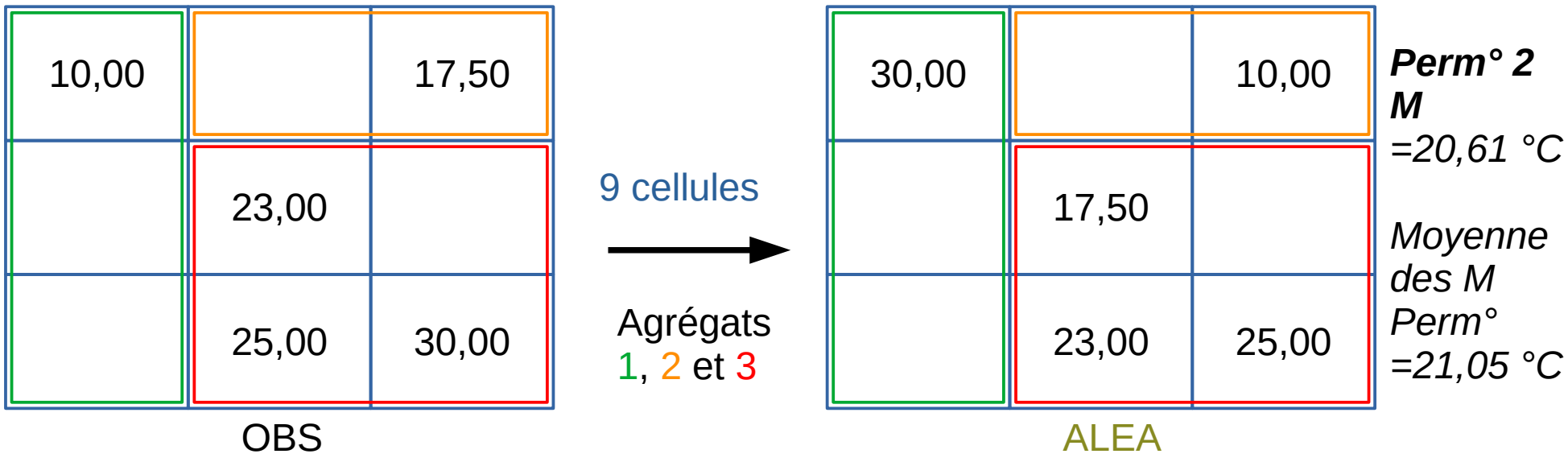
N=5
N_Agreg=3

Distribution spatiale des valeurs différentes

N=5
N_Agreg=3

Distribution aléatoire à topologie et distribution statistique équivalentes

Les mêmes parcelles sont affectées de mutations observées, tirées au hasard sans remise



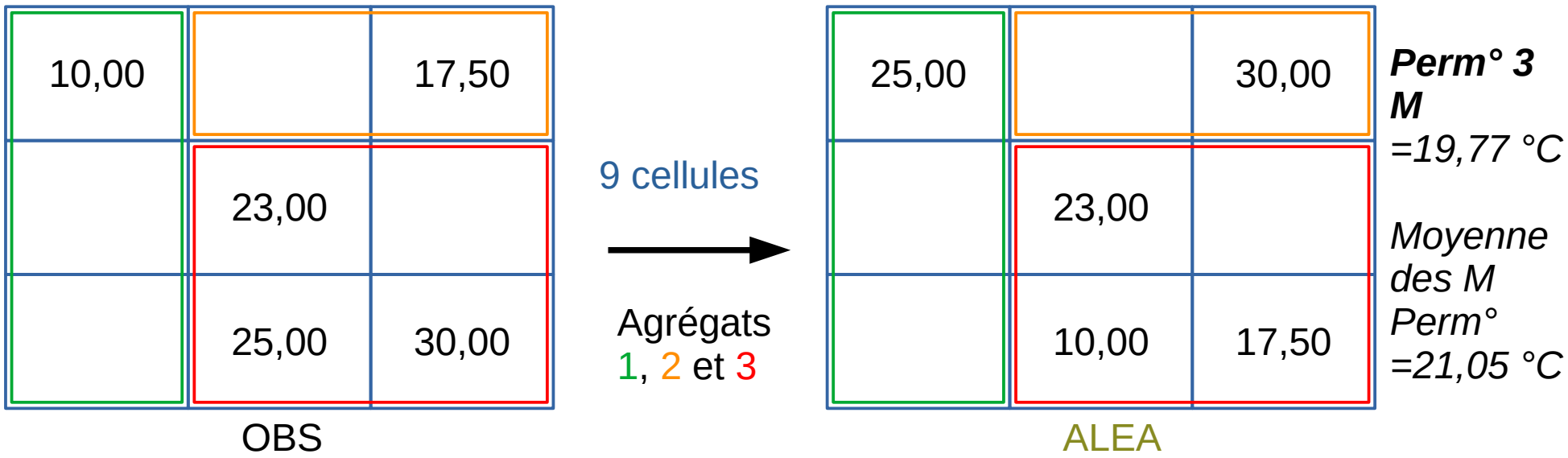
N=5
N_Agreg=3

Distribution spatiale des valeurs différentes

N=5
N_Agreg=3

Distribution aléatoire à topologie et distribution statistique équivalentes

Les mêmes parcelles sont affectées de mutations observées, tirées au hasard sans remise



N=5
N_Agreg=3

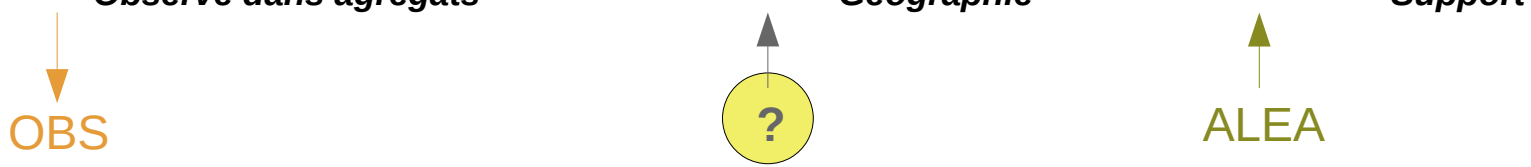
Distribution spatiale des valeurs différentes

N=5
N_Agreg=3

Effet du support spatial

Elimination de l'effet du support (DJ et al., 2012)

- Dans les calculs des indicateurs des valeurs observées via les permutations, le support spatial (topologie ST, distribution spatiale DG) et la distribution statistique des valeurs (DS) sont identiques
- Suite aux permutations, **la différence entre l'indicateur « observé » et l'indicateur « aléatoire » permet d'éliminer l'effet du support** car ce dernier agit dans les deux processus de calcul (DJ et al., 2012)
- Cet écart traduit la **part effective de la géographie dans la valeur mesurée**, puisque le processus de randomisation a supprimé toute autocorrélation spatiale
- $Indicateur_{Observé\ dans\ agrégats} = Indicateur_{Géographie} + Indicateur_{Support}$



Deux calculs de l'écart relatif ER

- Rapporté au cas aléatoire (%) : ***ER 1***

  100 x (Observation - Aléa) / Aléa

$$ER(\%) = 100 * \frac{|S_{obs} - S_{aléa}|}{S_{aléa}}$$

- Rapporté au cas observé (%) : ***ER2***

  100 x (Aléa - Observation) / Observation

- Dans les deux cas :

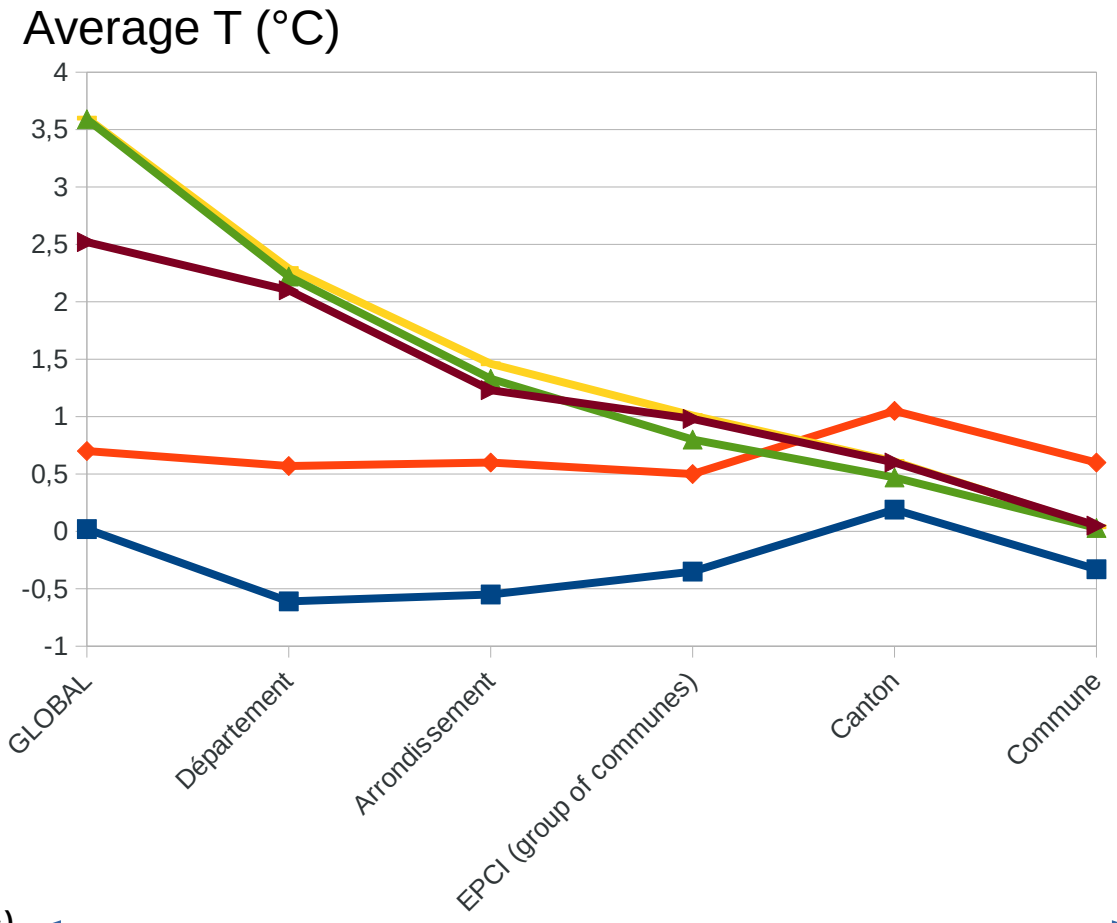
  => plus l'écart relatif absolu est élevé, plus la « part géographique de la mesure » est importante en rapport avec la référence (aléa vs observation)

  => Si ER → 0 : mêmes valeurs d'indicateur

  - *Les mutations des parcelles sont réparties aléatoirement sans structure, ni autocorrélation*
  - *Le seul effet est celui du support spatial, commun aux deux modes de calcul (observé vs aléatoire)*

Temperature measures through scales

Data: monthly minimal and maximal temperature (Meteo France Stations) in the PACA region

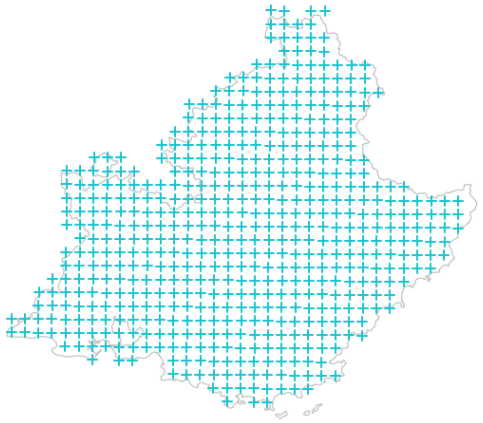


Case of **minimal temperature in January**

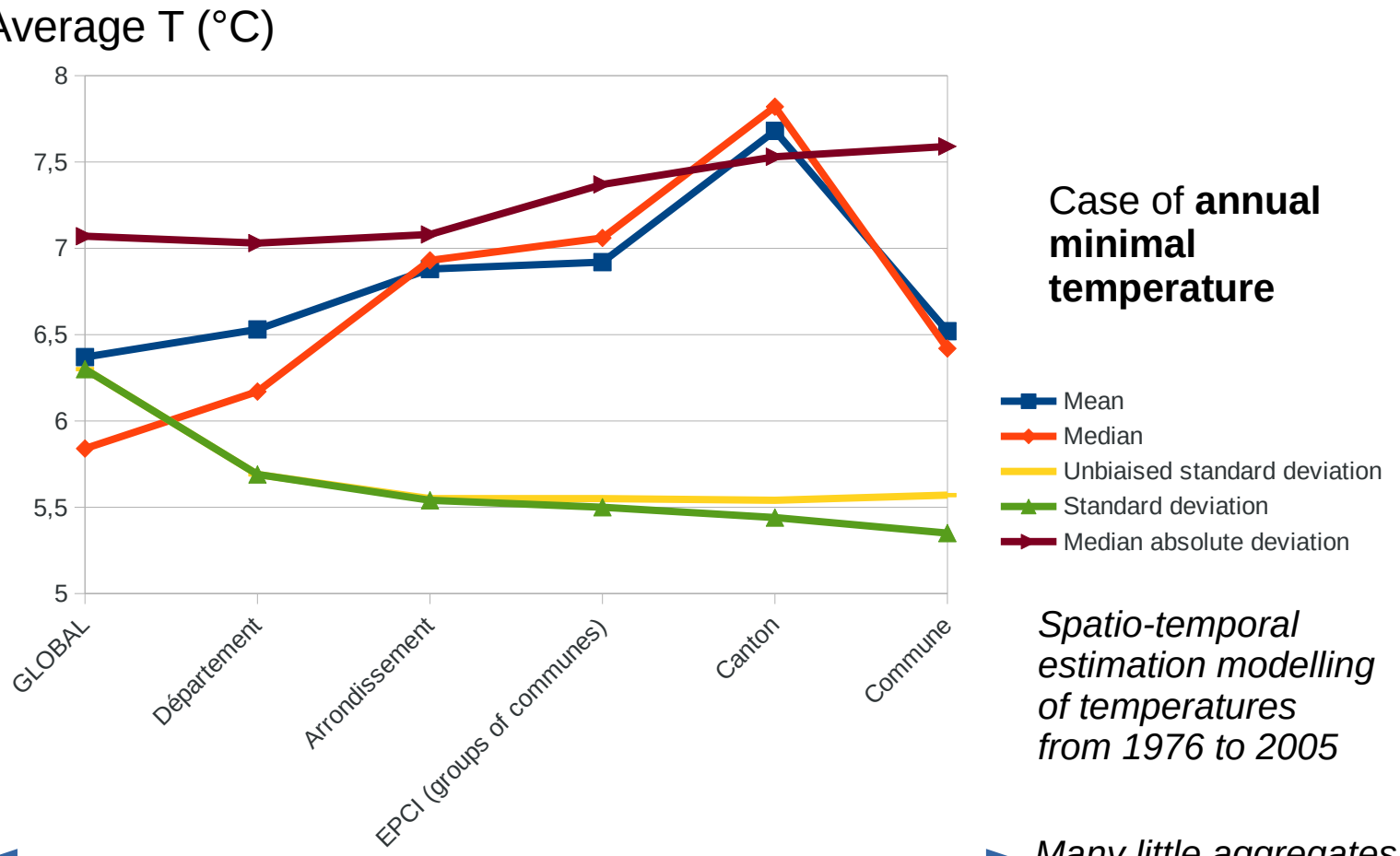
Temperatures observed from 1976 to 2005

Few large aggregate(s) ⟷ *Many little aggregates*

Temperature estimations through scales



Data: monthly extreme temperatures (ALADIN models from CRNM) in the PACA region



Case of **annual minimal temperature**

Spatio-temporal estimation modelling of temperatures from 1976 to 2005

Few large aggregate(s) ⟷ *Many little aggregates*

Temperature downscaling in the PACA region

- Topographic and geographical drivers
- ALADIN models
- Multiple linear regressions

Downscaling process

Altitude, slope, orientation, location distance to the sea, pixel enclosing, spatial roughness

Topographic and geographical drivers at 50 m of resolution

Minimal (Tn) and maximal (Tx) temperatures

Output from ALADIN model at 8 km of resolution

Data comparison

Identification of ALADIN model bias

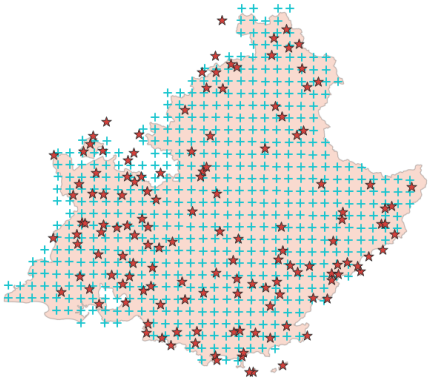
Data from meteorological stations

Multiple linear regression

Estimation of the model robustness

Regression validation and adjustment

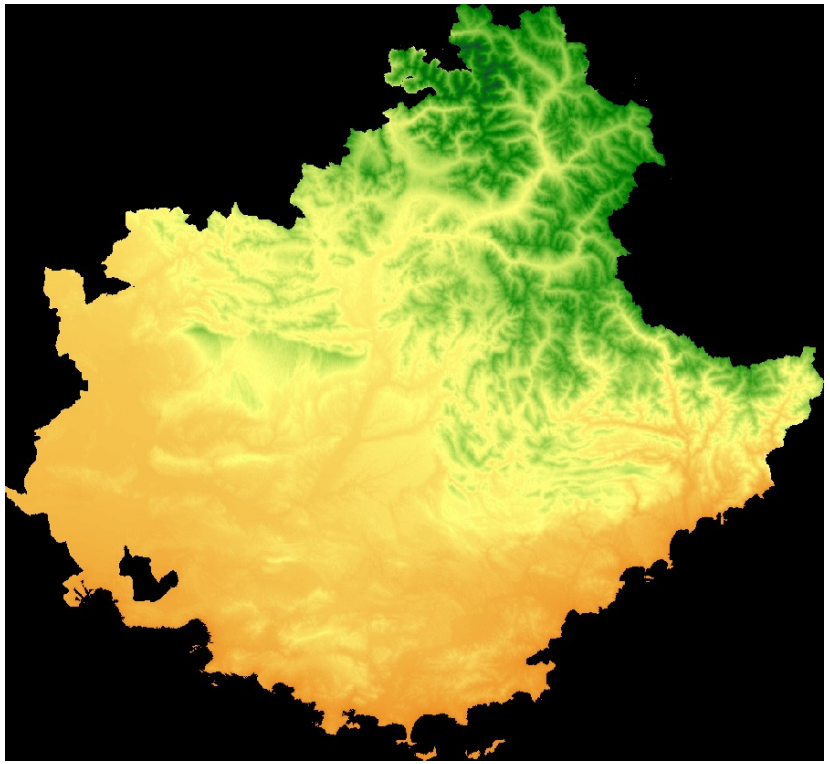
Maps of temperature at 50 m of résolution



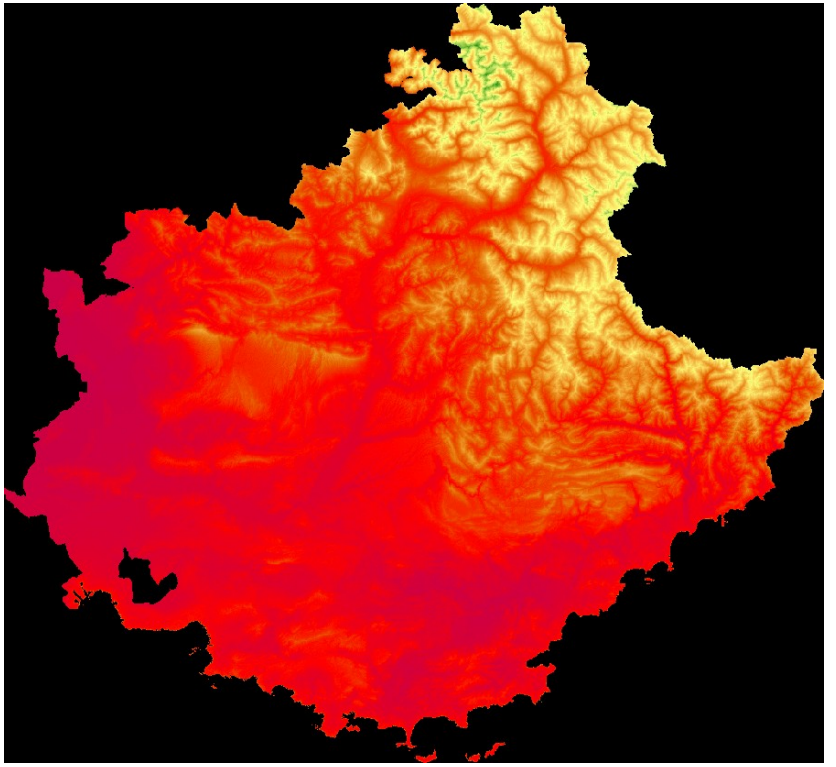
Alpes Maritimes

(Martin *et al.*, 2013
Vignal, 2018)

Maps of minimal and maximal Temperature



Minimal temperature in June



Maximal temperature in June

-10 -5 0 5 10 15 20 25 30 °C

Resolution : 50 meters

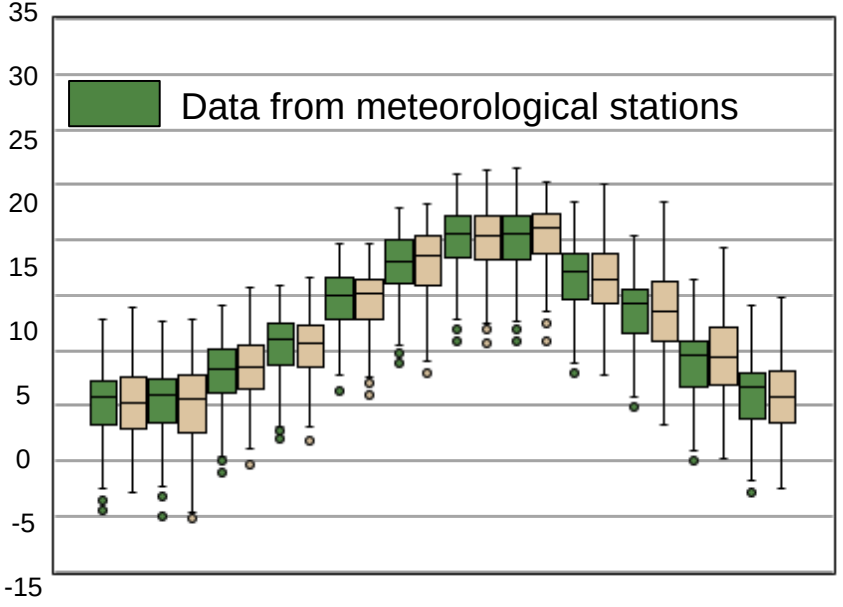
Vignal, 2018

Model *versus* observation

Minimal temperatures in June

	Intervals Tn
R^2	0.93 – **0.96** – 0.98
RMSE	0.53 – **0.66** – 0.85
Standard error	0.54 – **0.67** – 0.86

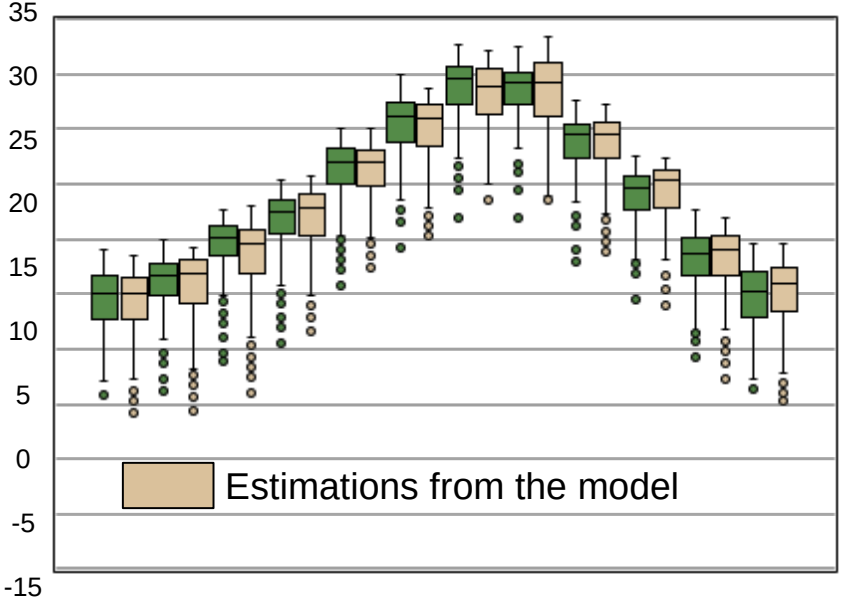
	Intervals Tn
Mean deviation	0.9 – **1.3** – 1.5
Standard deviation	0.7 – **1** – 1.1



Maximal temperatures in June

	Intervals Tx
R^2	0.98 – **0.98** – 0.99
RMSE	0.42 – **0.59** – 0.85
Standard error	0.43 – **0.60** – 0.86

	Intervals Tx
Mean deviation	0.7 – **0.9** – 1.3
Standard deviation	0.5 – **0.7** – 1



Vignal, 2018

Conclusion

- Be aware of the Change of (spatial) Support Problem:
  - Providing information at a given scale must be done with caution
  - Central value may fluctuate (mean, median)
  - Variance intra/inter-aggregates often changes for little samples

Carrots
Potatoes
Tomatoes
Leeks

- *Upscaling* - possible to know about uncertainty due to spatial support:
  - A priori no need of assumptions and drivers
  - High dependence to spatial support
- *Downscaling* - possible to reduce uncertainty:
  - Need of assumptions and drivers (expertise mandatory)
  - Spatial support has an impact at the first step

Objectifs du papier

Enjeux scientifiques

- Géo-climato :
  - Estimer l’effet éventuel des échelles sur le downscaling
  - Valider les approches de downscaling
  - Valider les approches d’upscaling
- Maths :
  - Formaliser, à travers les échelles, les liens entre variance intra et inter
  - Valider la méthode de calcul de l’écart relatif à travers les échelles

Méthodologie

- À l’échelle d’une partition donnée :
  - Chercher des données partinentes sur site bien connu
  - Réaliser les modèles de désagrégation
  - Conserver les indicateurs de qualité en fonction des densités de stations de référence
- Upscaling :
  - Cas 1 : partir des données des stations et agréger
  - Cas 2 : partir des estimations des modèles désagrés et ré-agréger
  - Conserver les indicateurs de qualité
- Comparaison des 2 approches

Positionnement global

Measuring the evolution of temperature in space is currently rather easy thanks to cheap and efficient sensors on a given location. However, from a set of measurement points, it becomes difficult to estimate or interpolate the real temperature values in every unknown locations. A response to this issue may be geostatistics. Denser the points, more efficient those methods, since they handle larger samples of data spread on restricted areas. Behind this relationship between space and measures samples is hidden the Modifiable Areal Unit Problem, that is highly related to the way these samples are spatially allocated and aggregated and consequently, how many data each sample counts. When processing downscaling at different nested spatial granularities, this change of data density can have a major effect on the quality of the estimation. That is what we suggest to explore, by recomposing and comparing the estimates following several random spatial partitions with the same sets of data. The scientists of ESPACE laboratory developed different methods and concepts about upscaling and dowscaling. Some previous works were done on upscaling using spatial segmentation algorithms on aerial images, on downscaling for assessing and monitoring temperature using statistics from ALADIN-climate models.

On two series of measures both in Southern France, relatively accurate in space (in Alpes Maritimes) and on a long term (in the Cévennes, up to 40 years of measures), we collected meteorological data. Those data a computed by downscaling models according to several spatial partitions randomly rebuilt. The results of the models are compared and discussed. The effect of the partition and the granularity is assessed and gives some room of confidence on the estimations provided by the models.

Papier commun (2023)

Data are computed by downscaling models according to several spatial partitions driven by several pertinent geographic and topographic factors. This generates an accurate and reliable information at the scale of 50 meters, by disaggregating climatological data from a scale of 8x8km. Using upscaling methods and an already published index to evaluate the change of support problem, we re-upscale these data up to the ALADIN grid (8x8 km) and compare the results with the raw data from the model.

The article highlights, while upscaling and downscaling, what are the main aspects of the relation between the data handled (here the temperature) and the spatial support effect of the partition, on both geographical and statistical points of view. It gives some room of confidence on the estimations provided by the models, after passing a rescaling process.

Data

- Data communes :
  - Alpes-Maritimes
  - températures : moyenne (min max moyenne) mensuelle
  - sur période de référence : 1981 => 2010 (ou autre ?)
  - Mêmes mailles amont / aval
- Variables explicatives :
  - Liste de variables possibles
  - Chacun choisit et fait ses modèles avec ses méthodes
  - Tester combinaisons des variables choisies (ensemble des combinaisons possibles à la base)
  - Même référence terrain (stations)

Modèle de désagrégation

- Validation données pour modélisation :
  - Vérifier conditions régression (normalité + VIF...)
  - Si pas de respect parfait : voir dans quelle mesure on peut le faire en gérant les problèmes (méthodes non paramétriques ?)
- Modèles :
  - Plusieurs modèles testés en parallèle au choix
  - En aveugle (X3 modèles)
  - Mêmes critères d'évaluation :
    - Critères de qualité : ajustement stat et aussi conditions (p-value)
    - Ajustement à la valeur mesurée dans les stations
- Validation des résultats des modèles et spatialisation multi-échelles (maille) :
  - Présélection indicateurs : moyenne, médiane, écart-type (histogramme des erreurs) : discussion
  - On sort un ou qqs modèles "clés"
  - Pour chaque étape d'agrégation, on compare et valide
  - On obtient une série de mailles (ou d'équations ?) à projeter sur la maille correspondante

Modèle d’agrégation

- Données sources
  - Source 1: semis de valeurs sur les stations
  - Sources 2 : nos grilles fines à agréger (obtenues par le modèle de désagrégation)
- Processus commun :
  - Agrégation emboîtée jusqu’à la maille la moins fine (la même que celle au départ de la désagrégation)
  - Obtention de tous les indicateurs à tous les niveaux d’échelle
- Comparaison des résultats des modèles
  - Différences entre données base downscaling et données reconstruites en upscaling
  - Répartition spatiale des différences (cartographie des résidus)

Impact of scales and sample size on temperature assessment within rescaling processes

Didier.josselin@univ-avignon.fr

Thank you for your attention

LMA
Laboratoire de Mathématiques d'Avignon

ESPACE