Département de Mathématiques d'Orsay

ÉCOLE POLYTECHNIQUE
UNIVERSITÉ PARIS-SACLAY

Species Abundance Monitoring with Crowdsourcing Data?

Christophe Giraud

Université Paris Sud
Ecole Polytechnique

Chaire MMB, décembre 2015

université PARIS-SACLAY

UNIVERSITÉ PARIS SUD
Comprendre le monde, construire l'avenir®

References

- C. Giraud, C. Calenge, C. Coron, R. Julliard. Capitalizing on opportunistic data for monitoring relative species abundances. To appear in Biometrics (2016).
- C. Giraud, R. Julliard, E. Porcher. Delimiting synchronous populations from monitoring data. Environmental and Ecological Statistics (2013).
- C. Calenge, J. Chadoeuf, C. Giraud, S. Huet, R. Julliard, P. Monestiez, J. Piffady, D. Pinaud, S. Ruette. The spatial distribution of Mustelidae in France. PLoS ONE (2015).

Species abundance monitoring

Different goals

- Atlas of presence (where can we find wolves in France?)
- Abundance estimation (how many wolves in Mercantour?)
- Investigation of spatial and/or temporal variations of abundance (temporal evolution of wolves in Mercantour compared to Savoie?)

Abundance versus relative abundance

Notation

N_{ij} = abundance of species i at time/location j

Relative abundance

For investigating spatio-temporal variations of abundance, we do not need to estimate absolute abundances but only relative abundances $N_{ij}/N_{ij'}$.

Our goal

In the following, our goal is only to estimate relative abundances

$$\widetilde{N}_{ij} = N_{ij}/N_{i1}.$$

We **do not try** to estimate abundances.

Data: institutional data

Institutional data

Data from scientists or environmental institution

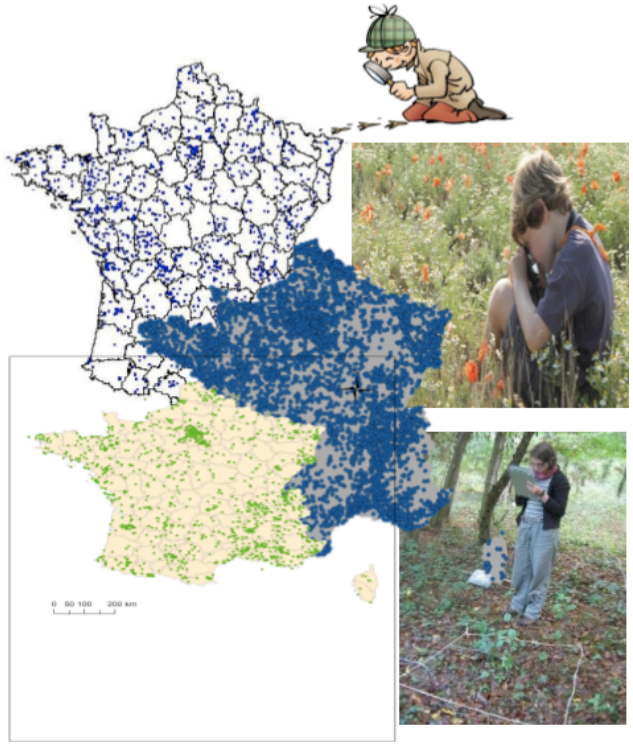
- universities
- conservation programs, national parks
- hunting management programs

Features

- High-quality data ☺
- Standardized protocols ☺
- Small coverage ☹

$\longrightarrow$ insufficient coverage for large scale relative abundance monitoring

Data: citizen participative data



Citizen data

Many different programs collect data, most of the time via a dedicated website. These programs are very heterogeneous.

Two important families of data

- Citizen science programs from some scientific institution
- Opportunistic data collection program (pure crowdsourcing)

Participative data: citizen science programs

Citizen science program

Institutional programs of data collection with usually:

- a standardized protocol
- some quality controls (of various nature)

Examples

- STOC data (MNHN): common birds (check list), stratified random sampling, 2 visits, 10 observation points of exactly 5 minutes within 4 hours after sunrise. Cooptation.
- SPIPOL data (MNHN): pollinator survey, 20min, with pictures of every insects (for identification), at any time. Open to anyone. Online identification from pictures (with cross-validation).

Participative data: citizen science programs

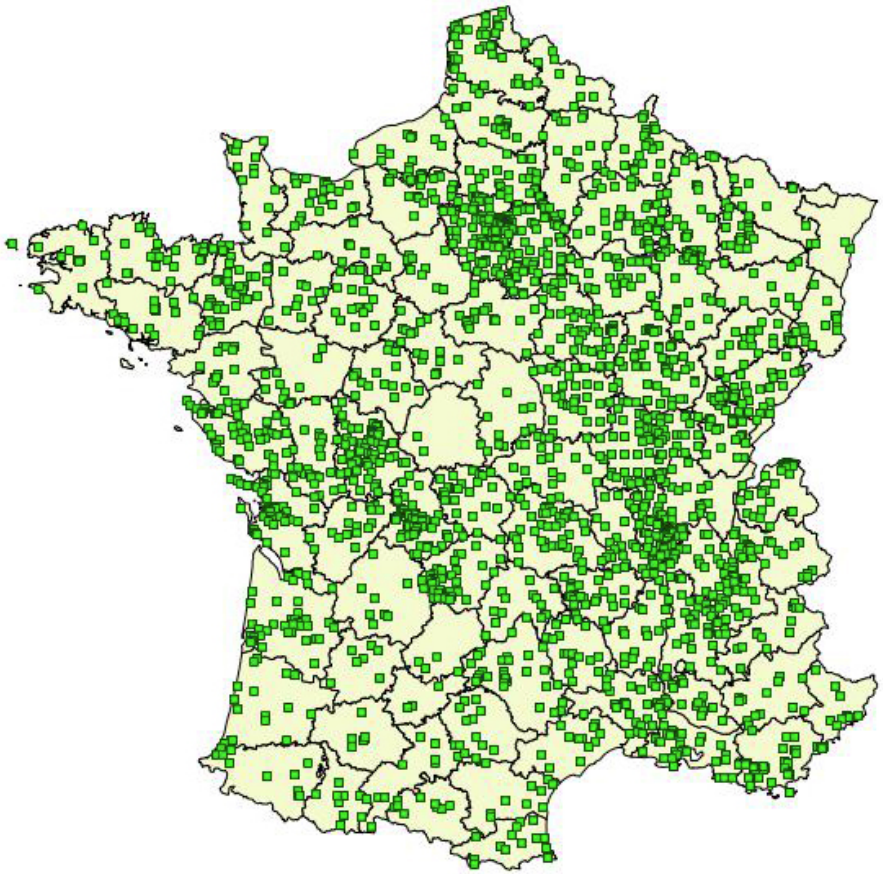
VIGIENATURE

Un réseau de citoyens qui fait avancer la science

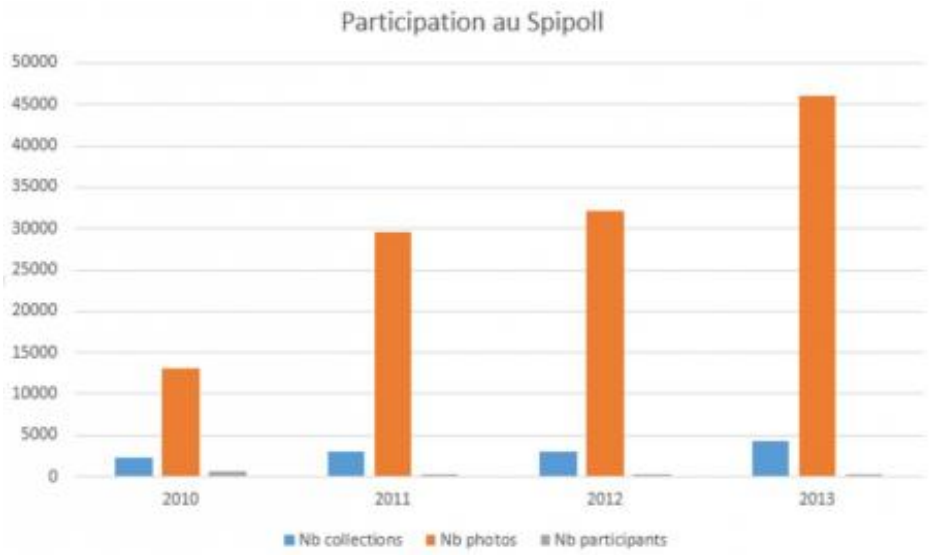
LE SITE

LE BLOG

Nos observatoires | POUR TOUS | NATURALISTES | GESTIONNAIRES D'ESPACES | AGRICULTEURS | ENSEIGNANTS | VIGIE-MANIP'



Participative data: citizen science programs



Features

- Quality controls ☺
- Standardized protocols ☺
- Medium coverage

$\longrightarrow$ insufficient coverage for some ecological investigations

Participative data: opportunistic data

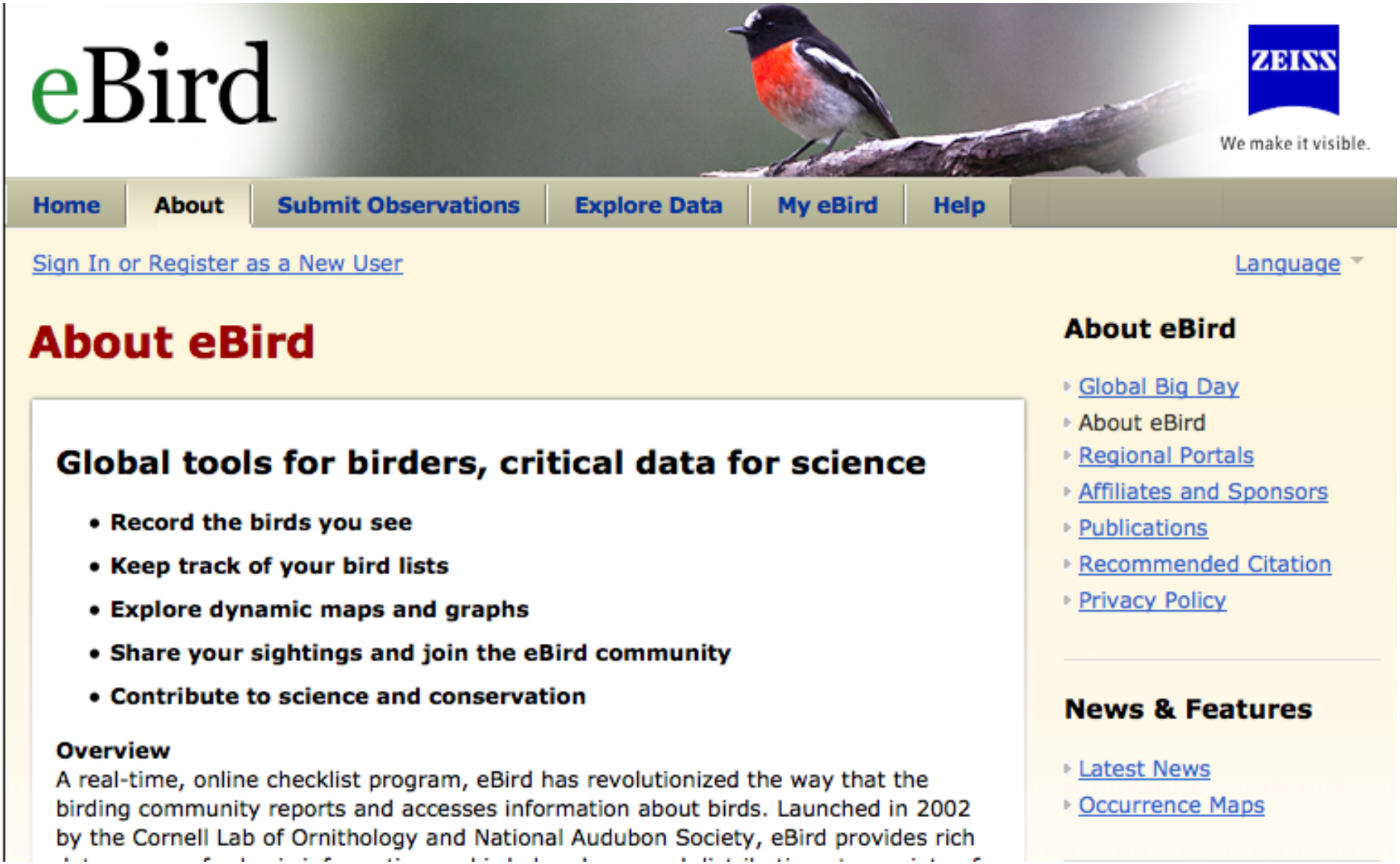
Opportunistic data

Data collections with usually:

- no protocol, no complete check-list,
- no quality controls,
- open to anyone,
- sometimes no scientific purpose (simply social sharing of observations).

Examples

- LPO (french ecological association for birds conservation): anyone can record his own observations after a birding session.
- eBird.org : similar in North America (mainly), with a temporal atlas of migrations
- Peau bleue : similar for divers


eBird
ZEISS
We make it visible.
Home
About
Submit Observations
Explore Data
My eBird
Help
Sign In or Register as a New User
Language
About eBird
Global tools for birders, critical data for science
• Record the birds you see
• Keep track of your bird lists
• Explore dynamic maps and graphs
• Share your sightings and join the eBird community
• Contribute to science and conservation
Overview
A real-time, online checklist program, eBird has revolutionized the way that the birding community reports and accesses information about birds. Launched in 2002 by the Cornell Lab of Ornithology and National Audubon Society, eBird provides rich
About eBird
Global Big Day
About eBird
Regional Portals
Affiliates and Sponsors
Publications
Recommended Citation
Privacy Policy
News & Features
Latest News
Occurrence Maps

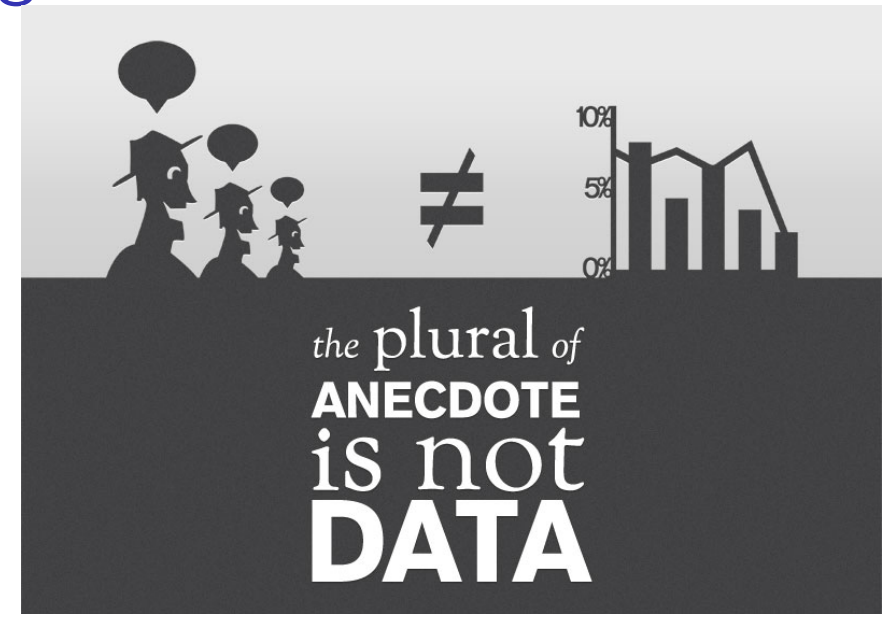
Miracle or Mirage?

MIRACLE !!

We have all the information: don't need stats anymore!

- Thousands of observers !
- Millions of counts !
- FREE !

Miracle or Mirage?



MIRAGE !!

The plural of anecdote is not data!

- no quality control
- partial reporting
- strong socio-geographic biases
- heterogeneity of the observers
- no information on the observational effort

Participative data: opportunistic data

Features (summary)

- Very heterogeneous ☹
- No protocols ☹
- Large coverage ☺

$\longrightarrow$ Can we do something with these data?

Im my point of view:

- we can certainly get informations from these data (ask Google...)
- but, *can we draw scientific conclusions from these data?*

May be yes, with a good knowledge of the data collection process.

Overview

Contents

Basic modeling of observations

Modeling of a count

The count X_{ij} for the species i at location j is

$$X_{ij} = \sum_{v_j \in \mathcal{V}_j} \sum_{a_{ij}=1}^{N_{ij}} Z_{a_{ij} v_j}, \quad \text{with} \quad Z_{a_{ij} v_j} \sim \mathcal{B}(p_{a_{ij} v_j}).$$

Approximative distribution

With Le Cam Inequality (hypotheses...)

$$\begin{aligned} \text{law}(X_{ij}) &\approx \text{Poisson}\left(N_{ij} \sum_{v_j \in \mathcal{V}_j} \bar{p}_{i v_j} \right), \quad \text{with} \quad \bar{p}_{i v_j} = \frac{1}{N_{ij}} \sum_{a_{ij}=1}^{N_{ij}} p_{a_{ij} v_j} \\ &\approx \text{Poisson}(N_{ij} O_{ij}). \end{aligned}$$

Observational bias

The observational bias

$$O_{ij} = \sum_{v_j \in \mathcal{V}_j} \bar{p}_{iv_j}$$

reflects the observational process

Abundance		**Effort**		**Errors**
Species abundance distribution	$\times$	Observers distribution Observational effort distribution	$\times$	Partial reporting Misidentification Misreporting

$\xrightarrow{\text{sampling}}$

Data

Animal counts

Main Assumption

Main assumption

We have the decomposition

$$O_{ij} \approx E_j P_i.$$

Interpretation

- E_j =function(prospecting effort in site j, weather conditions, etc)
- P_i = detection/reporting probability for the species i

Validity?

Can be justified when the sites j have homogeneous habitat type proportions.
If not? ... see later.

Identifiability

Model

$$\text{Count}(\text{species}{=}i, \text{site}{=}j) \sim \text{Poisson}(N_{ij}\, E_j\, P_i)$$

where

- N_{ij} = abundance of species i at site j
- E_j = prospecting effort at site j
- P_i = detection/reporting probability of species i

Identifiability issue

prospecting effort E_j: unknown for opportunistic data, even in relative scale $E_j/E_{j'}$.

$\longrightarrow$ we cannot have access to relative abundances $N_{ij}/N_{ij'}$ from the distribution. ☹

Rationale

What can we do?

- ~~Modeling E_j~~ (too complex – too sensitive)
- Combining opportunistic data with "effort standardized data"

"Effort standardized data" = data where we know (or can estimate) the ratios $E_j/E_{j'}$

Notation

Dataset labelling:

- $k = 0$: "effort standardized data"
- $k = 1$: opportunistic data

Rationale

Combining data sets (basic model)

$$\text{Count}(\text{species}{=}i, \text{site}{=}j, \text{data}{=}k) \sim \text{Poisson}(N_{ij}\, E_{jk}\, P_{ik})$$

with

- $E_{j0}/E_{j'0}$ known (institutional), $E_{j1}/E_{j'1}$ unknown (opportunistic)
- $E_{j1} \gg E_{j0}$ (in general)

☺ $2IJ$ observations for $IJ + 2I + J$ unknown parameters

Identifiability requires $I + 1$ additional constraints

A simple Generalized Linear Model

Generalized Linear Model

$$\text{Count}(\text{species}{=}i, \text{site}{=}j, \text{data}{=}k) \sim \text{Poisson}(\lambda_{ijk})$$
$$\text{with } \log(\lambda_{ijk}) = n_{ij} + e_{jk} + p_{ik}.$$

$\rightarrow$ we can estimate the relative abundance for each species

Does it make sense?

Gain of combining?

- In theory?
- In practice?

Theoretical gain of combining

Reduction of variance

- Single "standardized" dataset: with E_{j0} known

$$\text{variance}(\widehat{N}_{ij}^{(1)}) = \frac{N_{ij}}{P_{i0}E_{j0}}$$

- standardized+opportunistic datasets: with $E_{j1} \gg E_{j0}$

$$\text{variance}(\widehat{N}_{ij}^{(2)}) \approx \frac{N_{ij}}{P_{i0}E_{j0}} \times \underbrace{\frac{P_{i0}N_{ij}}{\sum_l P_{l0}N_{lj}}}_{\text{reduction factor}}$$

Remark: Combining gain is limited.

Theoretical gain of combining: explanation

A simple formula

If the ratios P_{i0}/P_{i1} are the same for all i, we have

$$\widehat{N}_{ij} = \frac{X_{ij0} + X_{ij1}}{\sum_l (X_{ij0} + X_{lj1})} \times \frac{\sum_l X_{lj0}}{E_{j0}},$$

and when $X_{ij1} \gg X_{ij0}$

$$\widehat{N}_{ij} \overset{X_{ij1} \gg X_{ij0}}{\approx} \frac{X_{ij1}}{\sum_l X_{lj1}} \times \frac{\sum_l X_{lj0}}{E_{j0}}.$$

Explanation

Roughly, dataset 1 provides a precise estimate of $N_{ij}/\sum_l P_{l0} N_{lj}$ and dataset 0 is used to estimate $\sum_l P_{l0} N_{lj}$

Another theoretical gain

Species not monitored in "standardized dataset"

- Correspond to $P_{i0} = 0$
- We can still estimate $\widehat{N}_{ij}$

Theoretical performance

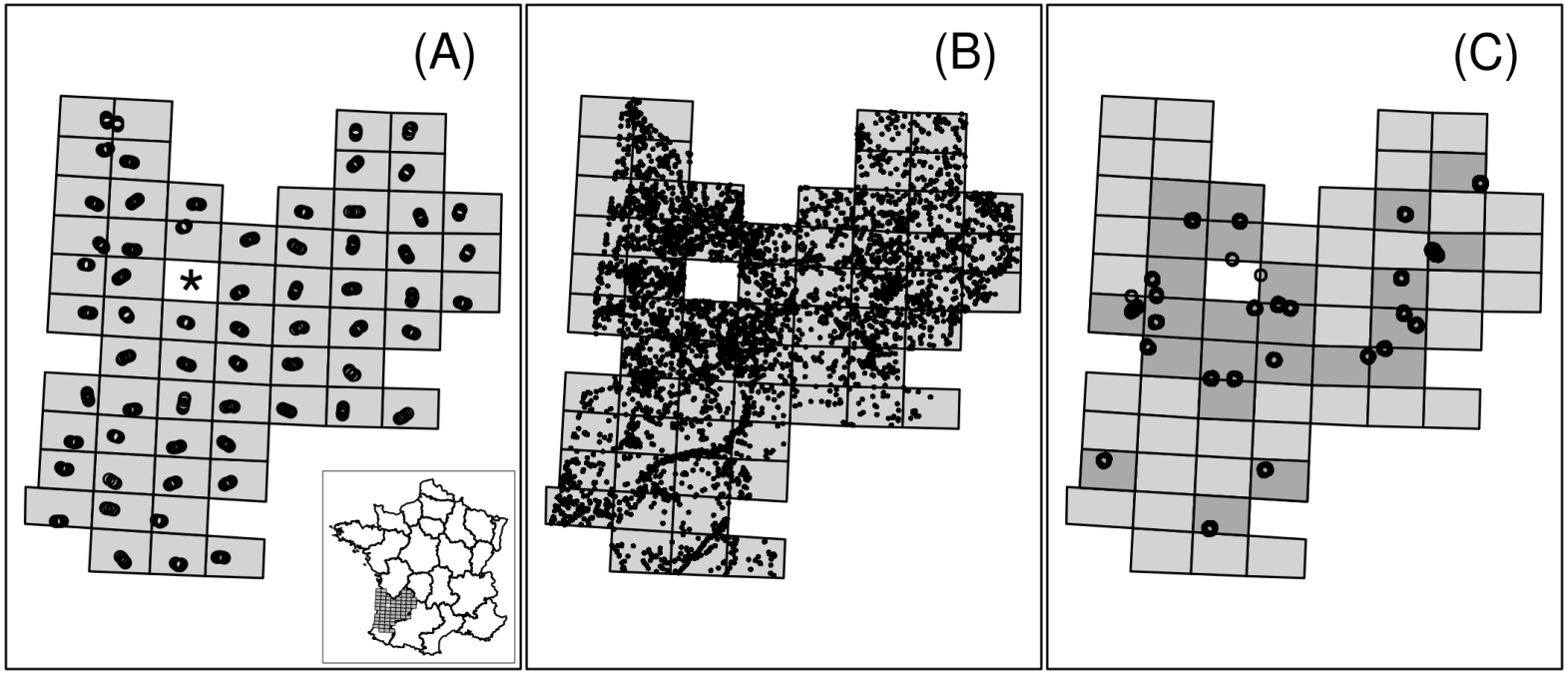
We have

$$\mathrm{var}(\widehat{N}_{ij}) \overset{E_{j1}\to\infty}{\sim} \frac{N_{ij}^2}{\sum_l P_{l0} N_{lj} E_{j0}}.$$

so in particular

$$\mathrm{var}(\widehat{N}_{ij}) \overset{E_{j1}\to\infty}{\sim} \mathrm{var}(\widehat{N}_{ij}^{0,\mathrm{imaginary}}) \times \frac{P_{i0}^{\mathrm{imaginary}} N_{ij}}{\sum_l P_{l0} N_{lj}}.$$

Numerical test



Dataset: birds in Aquitaine

Estimation:

(A) ACT dataset (from hunting management)

(B) LPO dataset (opportunistic)

Validation:

(C) STOC dataset (high-quality participative science)

Predictive power

Full ACT dataset

Data	ACT	ACT+LPO
ρ	0.27	0.55

Subsampled ACT dataset (1/18)

Data	ACT'	ACT'+LPO
ρ	0.06	0.54

$\longrightarrow$ clear gain of combining...

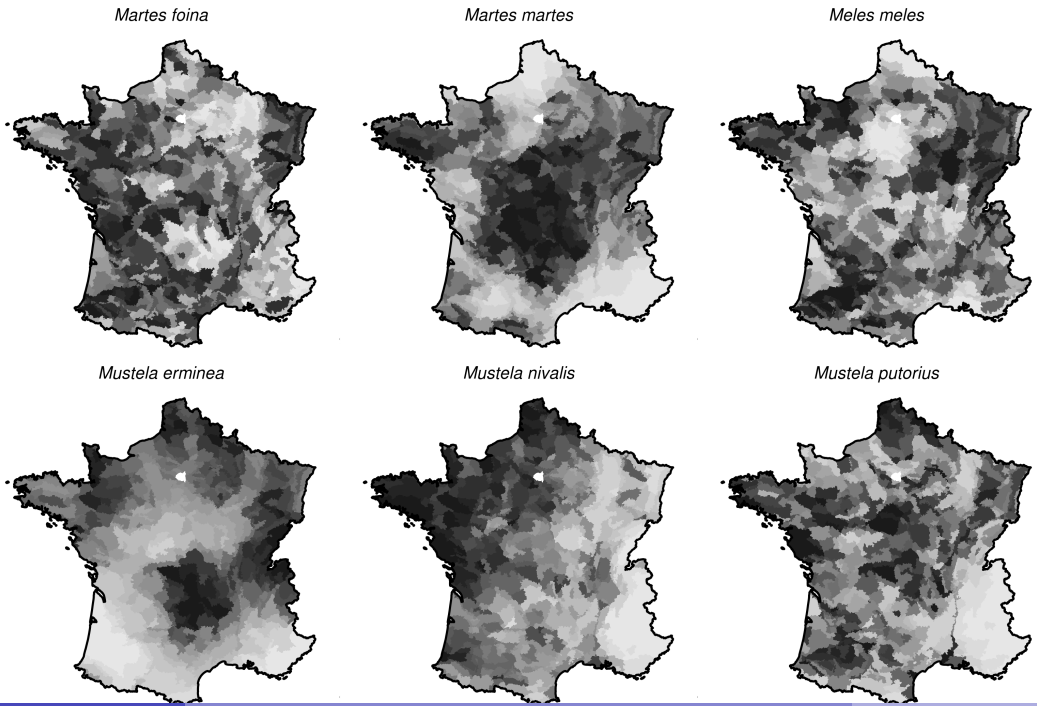
... despite the fact that the basic assumption $O_{ijk} = E_{jk} P_{ik}$ is not likely to be met (due to heterogeneous habitat type repartition)

Another example of application

Spatial distribution of Mustelidae

Data from a conservation institution (professional)

- proxy of the observational effort for part of the data
- but not for most of it



Another version of the combining idea (Fithian, Elith, Hastie & Keith, to appear in MEE)

Thinning model

- Species distribution: IPP(λ_i) with $\lambda_i(s) = \exp(\alpha_i + \langle \beta_i, x(s) \rangle)$
- Observations: IPP($\lambda_i b_{ik}$) with
  - $b_{i0}(s) = 1$ at locations where survey data are available, 0 else
  - $b_{i1}(s) = \exp(\gamma_i + \langle \delta, z(s) \rangle)$

Thought different (effort modeling, thinning), similar idea of "combining" survey data with opportunistic data.

A more realistic model

Stratification according to the habitat

Must take into account:

- Habitat type specificity of species: S_{ih}
- Habitat type bias of the observer: q_{hk}

$\longrightarrow$ requires a more complex modeling

$$X_{ick} \sim \text{Poisson}\left(N_{ij} E_{ck} P_{ik} \sum_{h \in c} \frac{q_{hk}}{\sum_{h' \in c} q_{h'k}} \frac{\alpha_h S_{ih} V_{hc}}{\sum_{h'} S_{ih'} V_{h'j}} \right).$$

Results: promising preliminary results.

Some opportunities for Ecology: an example

Delimiting synchronous populations

We want to identify areas where populations have the same temporal evolution (due to climatic forces, food avalaibility, etc).

Formalization

Find regions R such that

$$Z_{st} \sim \text{Poisson}(\exp(\theta_s + f(x_s, t))) \quad \text{with } f(x,t) \approx \sum_R \rho_R(t) \mathbf{1}_{x \in R}.$$

A tentative estimation procedure

Penalized negative log-likelihood

$$(\hat{\theta}, \hat{f}) \in \underset{\theta, f}{\operatorname{argmin}} \left\{ \mathcal{L}_Z(\theta, f) + \Omega(f) \right\},$$

with $\Omega(f)$ a convex penalty promoting solution with the shape

$$\hat{f}(x, t) = \sum_{j=1}^{J} \hat{\rho}_j(t) \mathbf{1}_{\hat{R}_j}(x).$$

Which penalty $\Omega(f)$?

Total variation norm

For $\mathcal{D}$ an open domain in $\mathbb{R}^d$ and $F : \mathcal{D} \to \mathbb{R}$.

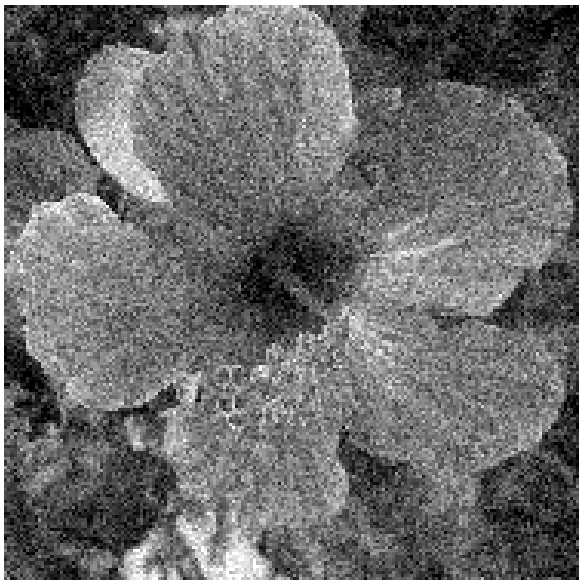
TV(F) norm

$$\begin{aligned} &\mathrm{TV}(F) \\ &= \sup\left\{ -\int_{\mathcal{D}} F(x)\mathrm{Div}(\phi(x))\,dx \;:\; \phi \in C_c^\infty(\mathcal{D}, \mathbb{R}^d) \text{ and } \|\phi\|_\infty \leq 1 \right\} \\ &= \int_{\mathcal{D}} \|\nabla F(x)\|\,dx \quad \text{if } F \in C^1. \end{aligned}$$

Reminder: $\mathrm{TV}(\mathbf{1}_R) = \mathrm{perimeter}\,(R)$

Example of use in image segmentation

Noisy image



TV regularized using λ=0.2

Synchronized Total Variation Norm

Our model

$$Z_{st} \sim \text{Poisson}(\exp(\theta_s + f(x_s, t))) \quad \text{with } f(x,t) \approx \sum_R \rho_R(t) \mathbf{1}_{x \in R}.$$

- **Similar :** at each time t the function $x \to f(x,t)$ is "block-constant"
- **New :** the blocks of $x \to f(x,t)$ must coincide at all time t.

Synchronized Total Variation Norm

STV norm

$$\begin{aligned} \mathrm{STV}(f) &= \sup_{\substack{\phi(.,t) \in C_c^\infty(\mathcal{D}, \mathbb{R}^d) \\ \| \sum_t \|\phi(.,t)\| \|_\infty \leq 1}} \left\{ -\sum_t \int_{\mathcal{D}} f(x,t) \mathrm{div}_x(\phi(x,t))\, dx \right\} \\ &= \int_{\mathcal{D}} \max_t \|\nabla_x f(x,t)\|\, dx, \qquad \text{when } f(.,t) \in C^1 \text{ for all } t. \end{aligned}$$

Properties

- $f \to \mathrm{STV}(f)$ is convex
- Minimizers of $\mathcal{L}_Z(\theta, f) + \alpha\, \mathrm{STV}(f)$ have the shape

$$\hat{f}(x,t) = \sum_j \hat{\rho}_{\widehat{R}_j}(t) \mathbf{1}_{x \in \widehat{R}_j}$$

Discretization issue

For images

straightforward discretization on a grid (discrete gradient)

For monitoring program

The observations are not spread on a grid $\longrightarrow$ the discretization is not straightforward.

Choice

Discretization of $\|\nabla_x f(x,t)\|$ with

$$\max_{u \in V(s)} |f(x_u,t) - f(x_s,t)|$$

where $V(s)$ is a neighborhood of s.

Estimation procedure

Estimator

We estime θ_s and $f(x_s, t)$ by $\hat{\theta}_s$ and $\hat{f}_{st}$ minimizing

$$\sum_{s,t} \left[e^{\theta_s + f_{st}} - Z_{st}(\theta_s + f_{st}) \right] + \alpha \sum_s \max_t \max_{u \in V(s)} |f_{st} - f_{ut}|$$

with $f_{s1} = 0$ for all s.

Optimization : with a primal-dual scheme (quite intensive)

Example: blackcap

Data

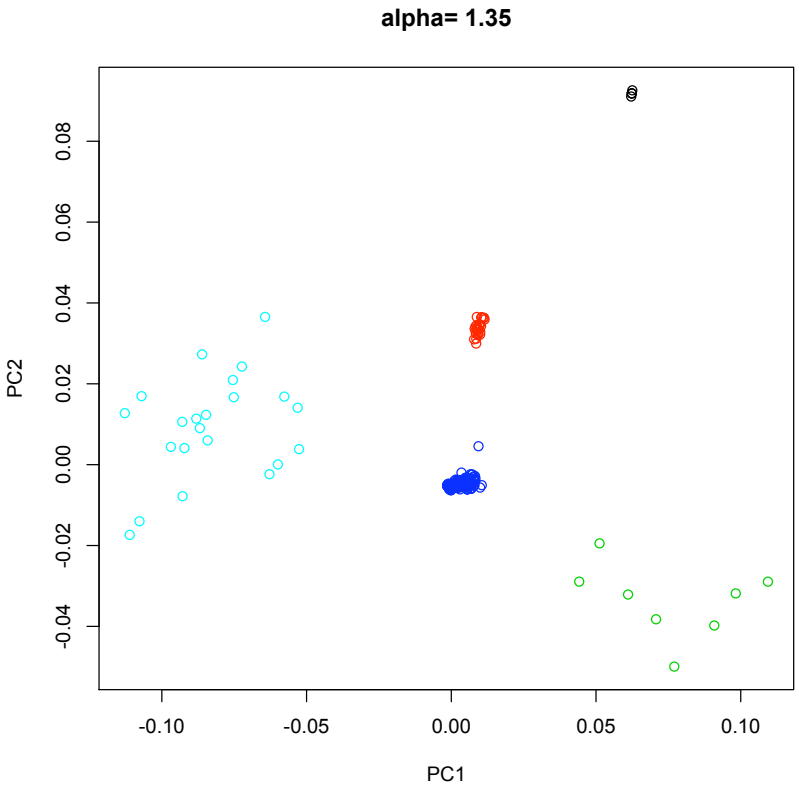
STOC data (only):

- high-quality "effort-standardized" data
- but many missing values
- from 2001 to 2009

Blackcap: 361 sites with at least 7 years of observations.

Results: estimated regions

Blackcap



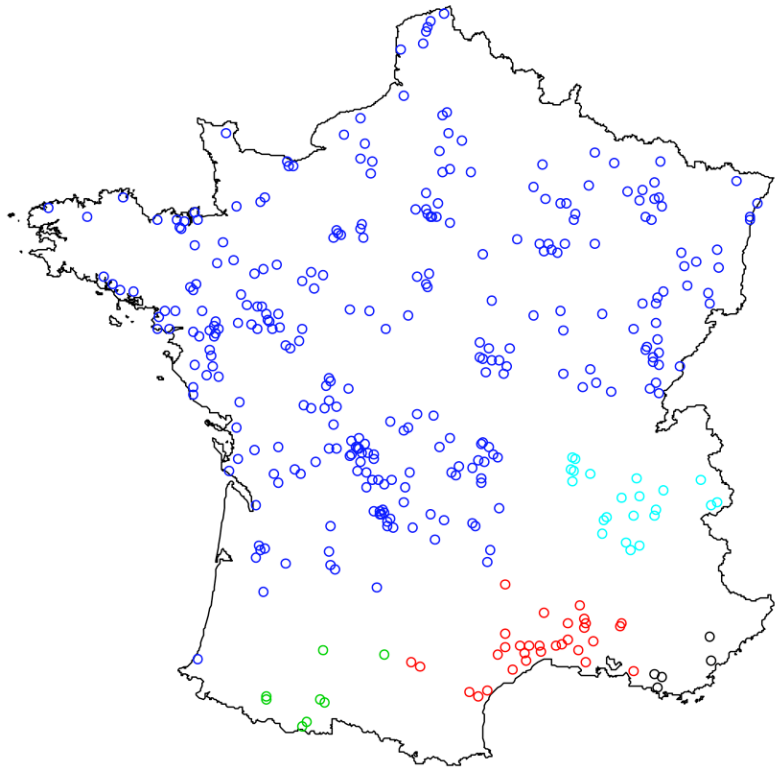
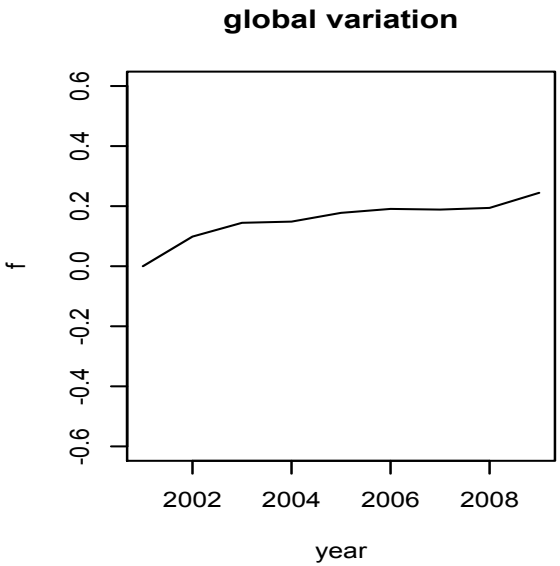
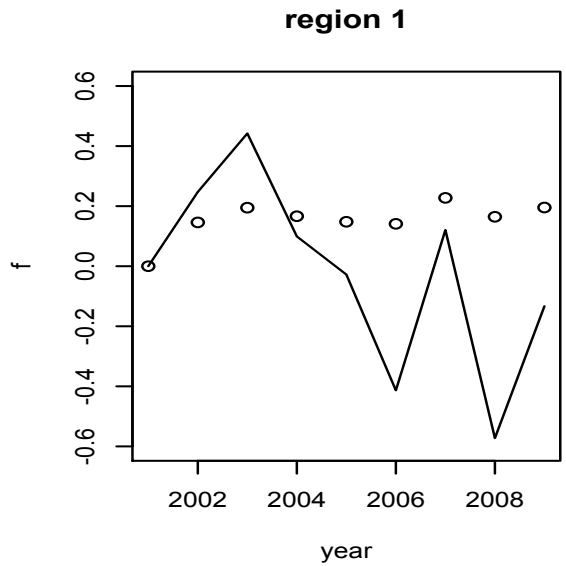
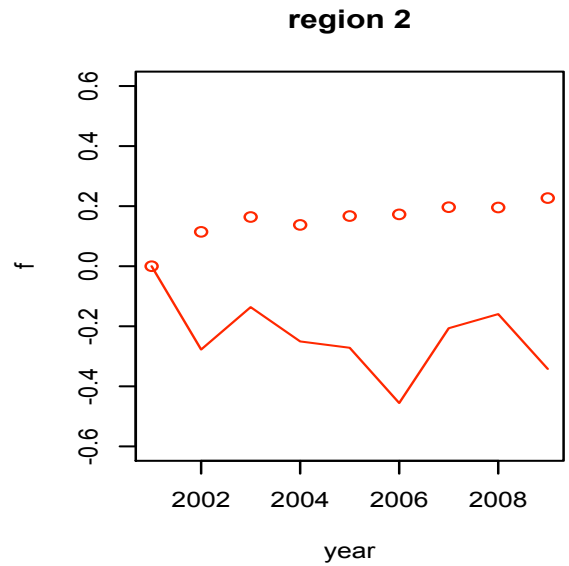


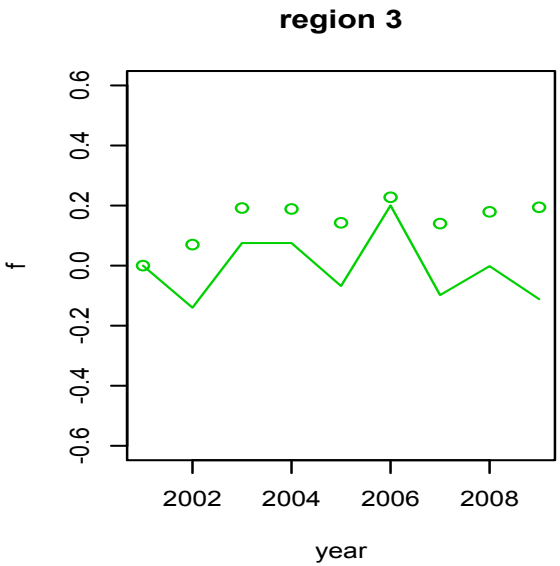
Figure: Left: PCA of $(\hat{f})$ and kmeans clustering. Right: regions selected by kmeans.

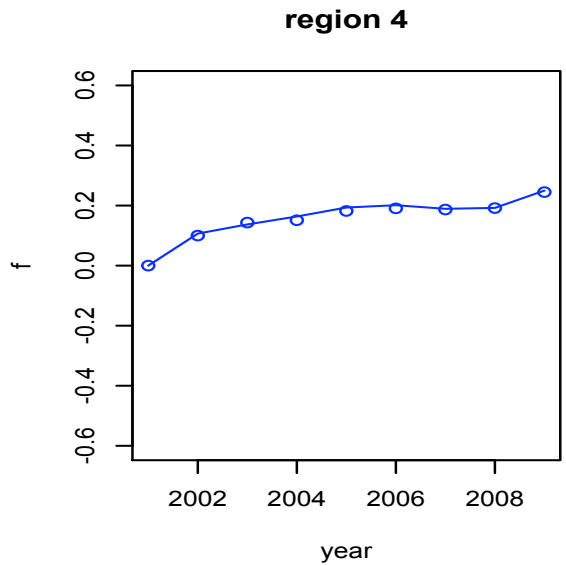
Results: temporal dynamics

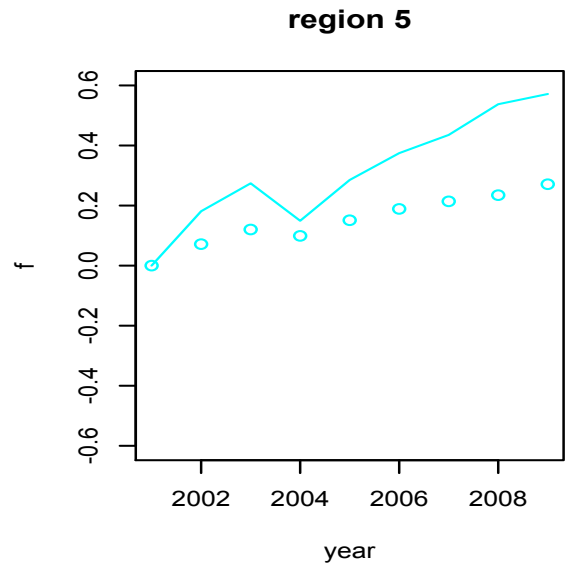












Results: discussion

The results fit the ecological knowledge ☺

- decline in mediterannean places
- increase in cold places

But, I think that

- insufficient coverage
- a better estimation procedure?

Perspective

Combining data

Combine standardized with opportunistic datasets

$$Z_{istk} \sim \text{Poisson}(\exp(\theta_{isk} + e_{stk} + f(x_s, t))) \quad \text{with } f(x,t) \approx \sum_R \rho_R(t) \mathbf{1}_{x \in R}.$$

and e_{st0} known up to (additive) constant.

Conclusion

- Opportunistic ecological data are massive but very heterogeneous.
- They can be useful for relative abundance monitoring.
- They could be crucial for investigating some fundamental questions in Ecology.