Time series analysis of absence/presence data in Ecology

Lionel Truquet (CREST-ENSAI, Rennes)

Research supported by the project **ECODEP**

Organization of the course

- Four sessions (January 24th and 31th, February 7th and 14th).
- The two first sessions are devoted to time series analysis of categorical data (theory and methods) as well as a general introduction to ergodic theory for time series analysis. **With me**.
- The two others sessions will be mostly devoted to the analysis of relative abundance data in ecology (theory for compositional data, i.e. data on the simplex). **With Guillaume Franchi**.

Outline for the two first sessions

- Motivation for modeling the dynamic of absence-presence data (**today on slides**).
- General introduction about ergodic theory for time series analysis (**today on tablet**).
- Theory for stationary autogressive categorical time series (**January 31th**).
- Inference methods (**January 31 th**).

Motivation

- In ecology, the study of **absence-presence of species** in an ecosystem is an important problem widely considered in the literature (Brotons et al. [Ecography, 2004], Mc Kenzie [Journal of wildlife management, 2005], Gormet et al. [Journal of applied Ecology, 2011]) .

- Such studies require to **explain or to forecast some binary vectors with coordinates** 0 **or** 1, depending if a given species is present or absent in a specific area.

- How to model the presence/absence data across the time and to identify possible patterns (attraction hypothesis between the species, influence of the environment, time dependencies...)?

- **Time series analysis of binary vectors is far from being well documented** if we target complex modeling (study of autoregressive processes, modeling the influence of exogenous regressors, spatio-temporal analysis if data are sampled at different sites).

Motivation through an example

- In Sebastián-González et al. [Proc. R. Soc. B, 2010], waterbird surveys are considered in a set of irrigation ponds.
- At each pond, the absence/presence data of 7 waterbirds were recorded during several years.
- Many covariates are available:
  - Fixed environmental and spatial covariates (pond area, presence or absence of shore/submerged/reed vegetation...).
  - Absence/presence of the same species at time $t-1$.
  - Absence/presence of other species at time t.
- The various covariates (time-varying and non time-varying) seem to have an impact on the dynamic.

Main questions for time series analysis

- How to develop autoregressive time-series models for binary data in which various types of covariates can be included ?
- What about probabilistic and statistical guarantees in the stationary case ?
- Use a general idea in econometrics: a categorical outcome can be seen as a discretization of a continuous one, i.e. if $Y_t \in \{0,1\}^k$ the absence/presence vector of k species at time t, for the species i,

$$Y_{i,t} = 1 \text{ if and only if } g_i\left(X_t\right) + \varepsilon_{it} > 0.$$

- X_t **is a vector of covariates (exogenous)** available at time t and at previous times. ε_t **is a noise component** (independent from X_t). X_t can contain lags values of the response, i.e. $Y_{t-1}, Y_{t-2}, \ldots$.
- Simplest choice for g: g_i linear function, i.e.

$$g\left(X_t\right) = d + BX_t.$$

Multivariate probit/logistic model in the static case

- At a given pond, let $Y_t \in \{0,1\}^k$ the absence/presence vector of k species at time t.

$$Y_{it} = \mathbb{1}_{\lambda_{it}+\varepsilon_{it}>0}, \quad \lambda_t = g(X_t) = d + BX_t.$$

- ε_t has a multivariate probability distribution with specified margins.

- A **multivariate Probit** model is obtained when ε_t is a Gaussian vector with mean 0 and correlation matrix R. This version is widely used in Econometrics (Chib and Greenberg [Biometrica, 1998]).

- A **multivariate logistic** model is obtained when $\varepsilon_{it} = F^{-1} \circ \Phi\left(V_{it}\right)$ where Φ is the c.d.f. of the standard Gaussian distribution, F is the logistic c.d.f., i.e. $F(x) = (1+\exp(-x))^{-1}$ and $V_{\cdot,t} \sim \mathcal{N}_k(0,R)$. See O'Brien and Dunson [Biometrics, 2004].

Model interpretation

- The models writes as

$$Y_{it} = \mathbb{1}_{\lambda_{it}+\varepsilon_{it}>0}, \quad \lambda_t = g\left(Y_{t-1}, \ldots, Y_{t-p}, X_t\right) = d + \sum_{\ell=1}^{p} A_\ell Y_{t-\ell} + BX_t.$$

- The sign of $A_j(i, i')$ indicates if the presence of species i' in the past has a positive or negative influence on the presence of species i at time t (**graphical interpretation**).

- The signs of the entries of the B matrix indicates a positive or negative impact of the covariates on the presence of species at time t.

- If $\varepsilon_t \sim \mathcal{N}_k(0, R)$, $R_{i,i'}$ indicates a positive or negative association between the presence of species i and i' at the same time t. For instance, Slepian's lemma guarantees that **Cov** $\left(Y_{i,t}, Y_{i',t} | X_t, Y_{t-1}, \ldots\right)$ **is increasing w.r.t.** $R_{i,i'}$ **and has the sign as** $R_{i,i'}$.

References for the two first sessions

- Debaly, Z.M. and Truquet, L., *Iterations of dependent random maps and exogeneity in nonlinear dynamics.* Econometric Theory, 2021, **37(6)**, 1135–1172.
- Samorodnitsky, G., *Stochastic processes and long range dependence.* 2016, **26**, Springer. **Chapter 2 for the ergodic theory**.
- Sebastian-Gonzalez, E., Sánchez-Zapata, J. A., Botella, F., and Ovaskainen, O. *Testing the heterospecific attraction hypothesis with time-series data on species co-occurrence.* 2010, Proceedings of the Royal Society B: Biological Sciences, 277(1696), 2983-2990.
- Truquet, L. (2021) *Strong mixing properties of discrete-valued time series with exogenous covariates.* arXiv:2112.03121.
- Truquet, L. (2022) *Theory and inference for mutivariate autoregressive binary models with an application to absence/presence data in Ecology.* **Not yet available, I hope to finish it very soon**.
- Wilkinson, D.P., Golding, N. Guillera-Arroita, G., Tingley, R. and McCarthy, M.A. *A comparison of joint species distribution models for presence–absence data.* Methods in Ecology and Evolution, 2019, **10(2)**, 198–211.