Modélisation stochastique et spectrale de l'occupation du sol

J.-F. Mari[1] et O. Horn[2]

[1]Université de Lorraine, Loria, UMR 7503

[2]Université de Lorraine, LCOMS, ISEA

SFC 2019, Nancy

Analyse spectrale d'une série temporelle d'items

Motivations

- L'activité humaine est source de données temporelles de différentes natures :

  données numériques : le prix du pain, les prix de l'essence à la pompe, le nombre de touristes à l'année ...

  données catégorielles : les majorités parlementaires, les occupations annuelles d'une parcelle agricole : *Blé*, *Orge*, *Colza* ...
- L'observation de comportements périodiques est une étape intéressante pour l'extraction de connaissances ;
- Permet une description plus concise des données.

Analyse d'une série temporelle

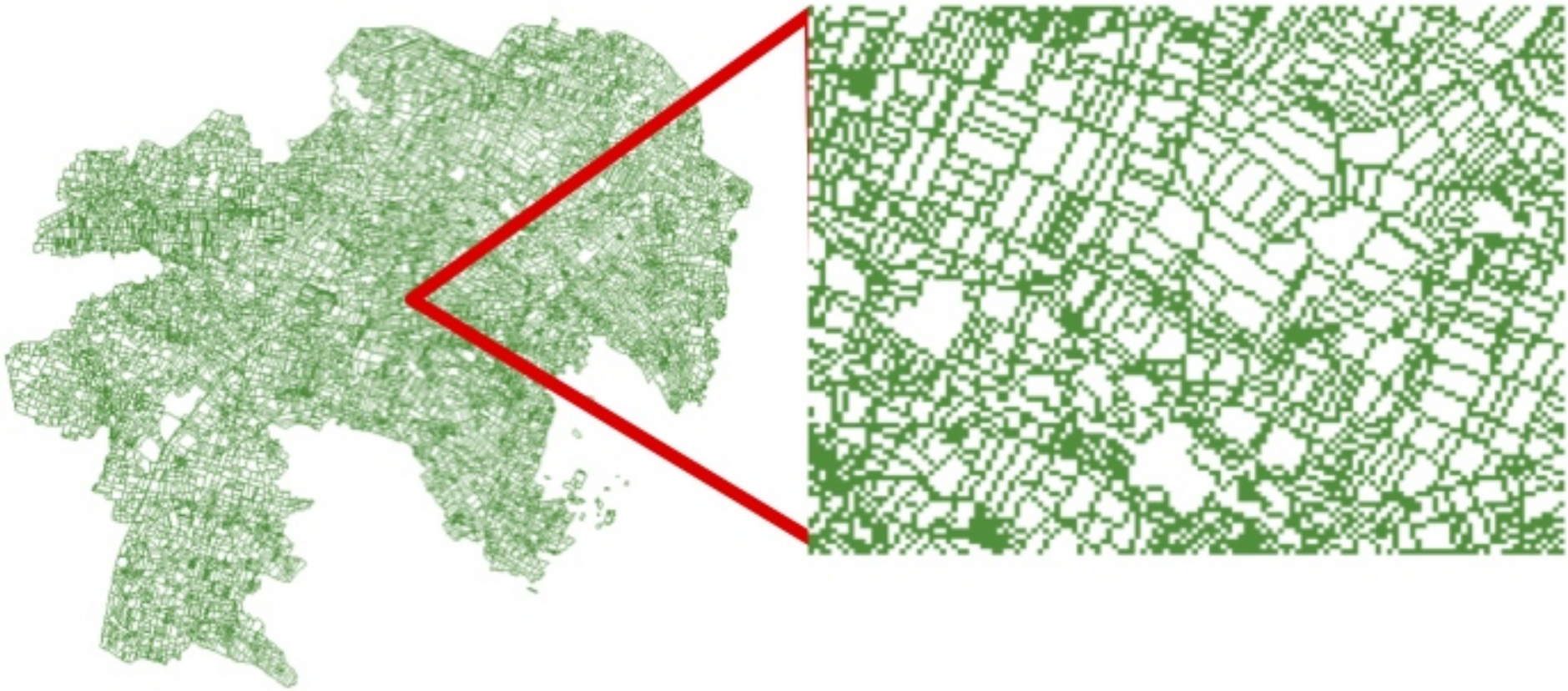
état de l'art

Méthodes combinatoires pour des séries temporelles d'items [Galbrun *et al.*2018, Elfeky *et al.*2005, Tatavarty *et al.*2007] ;

Analyse spectrale des fonctions d'auto corrélation sur des signaux continus échantillonnés : reconnaissance de la parole, recherche de périodes dans les traces de parcours d'animaux . . . [Vlachos *et al.*2005] et [Li *et al.*2012]

Notre approche Représentation tensorielle des donnnées catégorielles équivalente aux échantillons d'un signal continu pour effectuer une analyse spectrale (FFT) du signal d'auto covariance produit.

La mosaïque parcellaire et ses occupations



Definition

- Issues chaque année des relevés sur le terrain;
- Délimitées par des limites physiques;
- Ne contiennent qu'une occupation : colza, blé, orge, forêt, ...

Les données enquêtées

	1	2	$\cdots$	t	$\cdots$	$t+\tau$	$\cdots$
	$\vdots$			$\vdots$		$\vdots$	
$parcelle_i$	*blé*	*orge*	*colza*	*blé*	*orge*	*colza*	$\cdots$
	$\vdots$			$\vdots$		$\vdots$	
$parcelle_k$	*tournesol*	*blé*	*colza*	*blé*	$\cdots$		
	$\vdots$			$\vdots$		$\vdots$	

Propriétés des séries

séries stationnaires Les moyennes d'ensemble sont indépendantes du temps. Exemple : La distribution de culture sur l'ensemble du territoire est identique quelque soit l'instant choisi pour la calculer ;

séries ergodiques Les moyennes temporelles sont indépendantes des réalisations. Exemple : la distribution de culture sur la période d'étude est la même quelque soit la parcelle (donc de son propriétaire).

Quand l'ensemble des opportunités ou contraintes (économiques, climatiques, . . .) imposées aux cultivateurs ne changent pas, les séries temporelles d'occupations de parcelles agricoles sont ergodiques et stationnaires à petite échelle temporelle et spatiale.

Le modèle stochastique : signaux aléatoires

Definition

- $\mathcal{E} = \{e_1, e_2, \ldots e_K\}$ les K différentes occupations ;
- $x_1, x_2, \ldots x_T$ la série temporelle des occupations d'une parcelle ω pendant T années ;
- $X_t(\omega)$ la série temporelle de variables aléatoires dont les réalisations sont les x_t sur une parcelle ω ;
- On suppose $X_t(\omega)$ stationnaire et ergodique ;
- Le relevé d'enquêtes sur un territoire est vu comme un échantillonage de $X_t(\omega)$.

Les moments croisés

second moment croisé

$$\begin{aligned} C_X(\tau) &= E\left[X_t(\omega)X^*_{t+\tau}(\omega)\right] \\ &= \frac{1}{T}\sum_t x_t x^*_{t+\tau} Prob(x_t, x_{t+\tau}) \end{aligned} \tag{1}$$

x^* représente la transposé du vecteur x.

La fonction d'auto covariance

La matrice $K \times K$ définie en centrant le moment croisé.

$$R_{XX}(\tau) = E\left[X_t(\omega)X^*_{t+\tau}(\omega)\right] - E^2\left[X(\omega)\right] \tag{2}$$

Représentation des données sous forme de tenseurs

Le vecteur $\delta^i \in \mathcal{R}^K$, $\delta^i = \begin{bmatrix} 0 \\ \dots \\ 1 \\ \dots \\ 0 \end{bmatrix} \leftarrow i$ représente l'item e_i.

$Prob(\delta_t^i, \delta_{t+\tau}^j) = Prob(e_i \text{ au temps } t, e_j \text{ au temps } t+\tau)$

Dans un processus stationnaire et ergodique, les moyennes temporelles et d'ensembles sont constantes.
Le terme général (i, j) est égal à :

$$R_{XX}(\tau)^{(i,j)} = \frac{1}{T} \sum_t Prob(\delta_t^i, \delta_{t+\tau}^j) - E^i E^j \quad (3)$$

E^i est la composante i du vecteur $E[X(\omega)]$.

Estimation des moments des séries multidimensionnelles

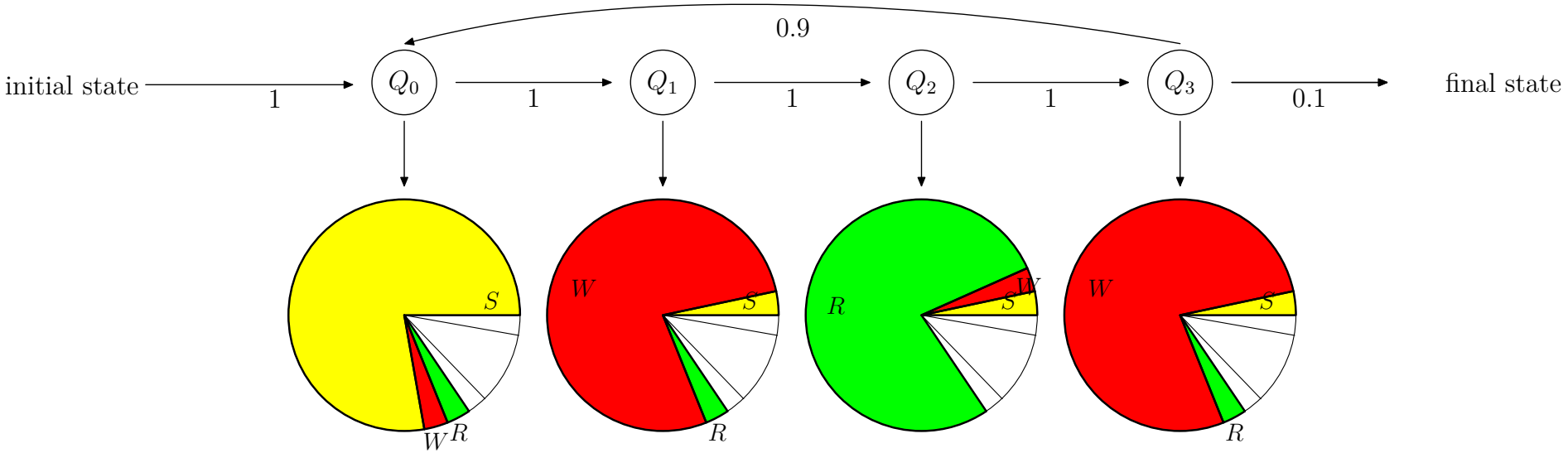
$Prob(\delta_t^i, \delta_{t+\tau)}^j)$ est estimé à partir de la fréquence des items e_i au temps t et e_j au temps $t+\tau$ dans le corpus d'enquêtes.

	1	2	$\cdots$	t	$\cdots$	$t+\tau$	$\cdots$
	$\vdots$			$\vdots$		$\vdots$	
ω	$\cdots$			e_i		e_j	$\cdots$
	$\vdots$			$\vdots$		$\vdots$	

Sur la même parcelle ω, on observe la culture e_i au temps t et e_j au temps $t+\tau$.

Construction de séries temporelles artificielles

Utilisation d'un modèle de Markov caché



`HMM` pour simuler des répétitions du patron [S-W-R-W].

Les Q_i sont les états du `HMM`.

Les séries sont bruitées. Les camemberts représentent les probabilités utilisées pour générer les différents items à chaque état.

Construction de 50 séquences de longeur 32 dans lesquelles on cherchera une période de 4 pour les items S et R et de 2 pour l'item W

Objectifs de l'analyse

On obtient un ensemble de séries temporelles de la forme :
S, W, R, W, S, W, R, W ...
dans laquelle des substitutions de symboles ont été effectuées.

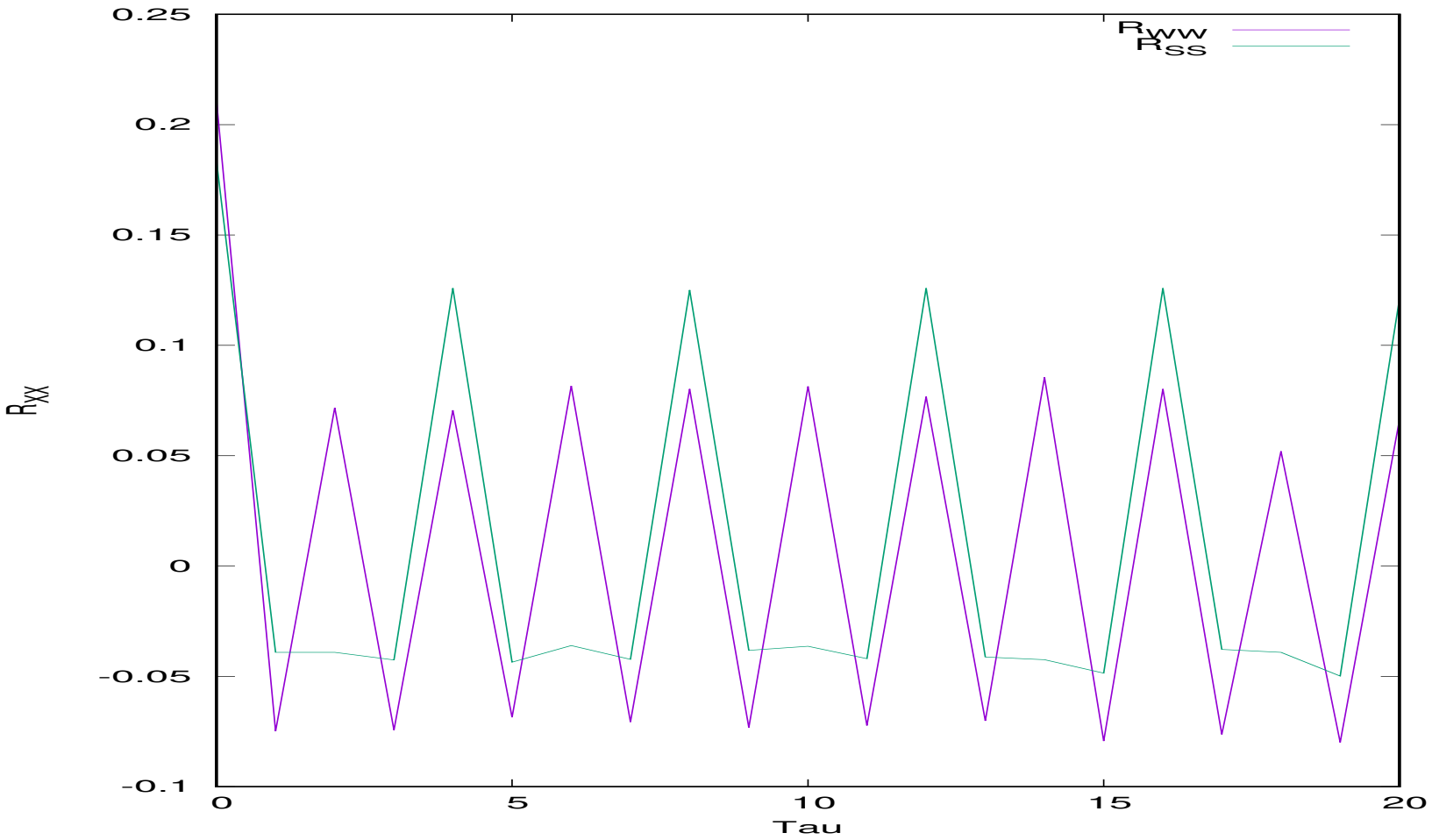
Recherche de période pour l'item W : période 2 pour **W** (ou temps de retour pour les agronomes) grâce à une FFT;

Recherche de période pour l'item S et R : période 4 pour les items **S** et **R** grâce à une FFT ;

Recherche d'itemsets : on doit trouver les itemsets **[SW]** et **[S ? R]** par l'étude de la fonction d'auto covariance croisée comme fonction du décalage (*lag*).

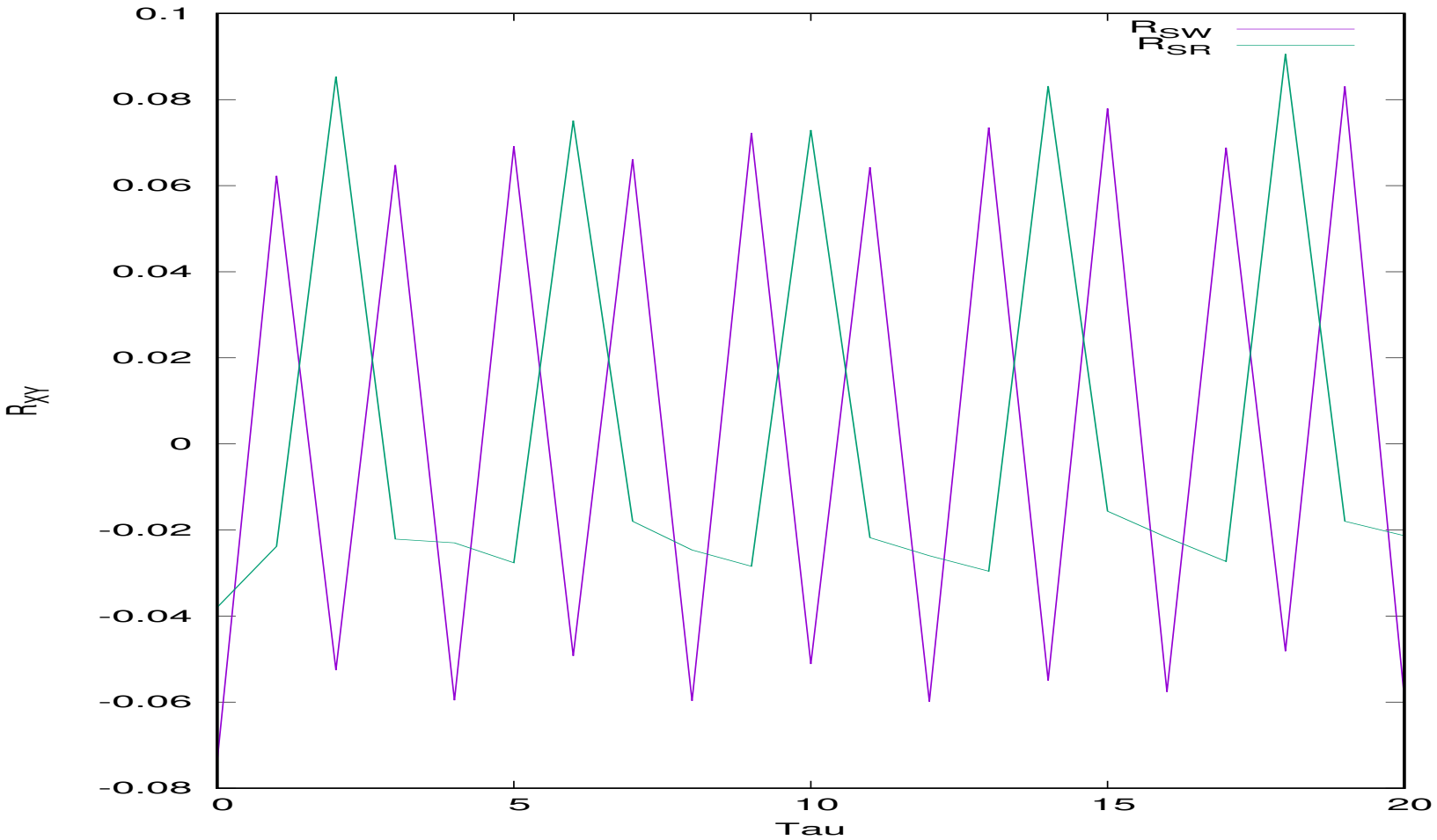
Signal d'auto covariance en fonction du temps sur des données artificielles

Mise en évidence des périodes de W et S



Signal de covariance croisée sur des données artificielles

Décalages (lags) entre S-W (=1) et S-R (=2)



On retrouve les itemsets [S, W] et [S ? R]

Analyse spectrale

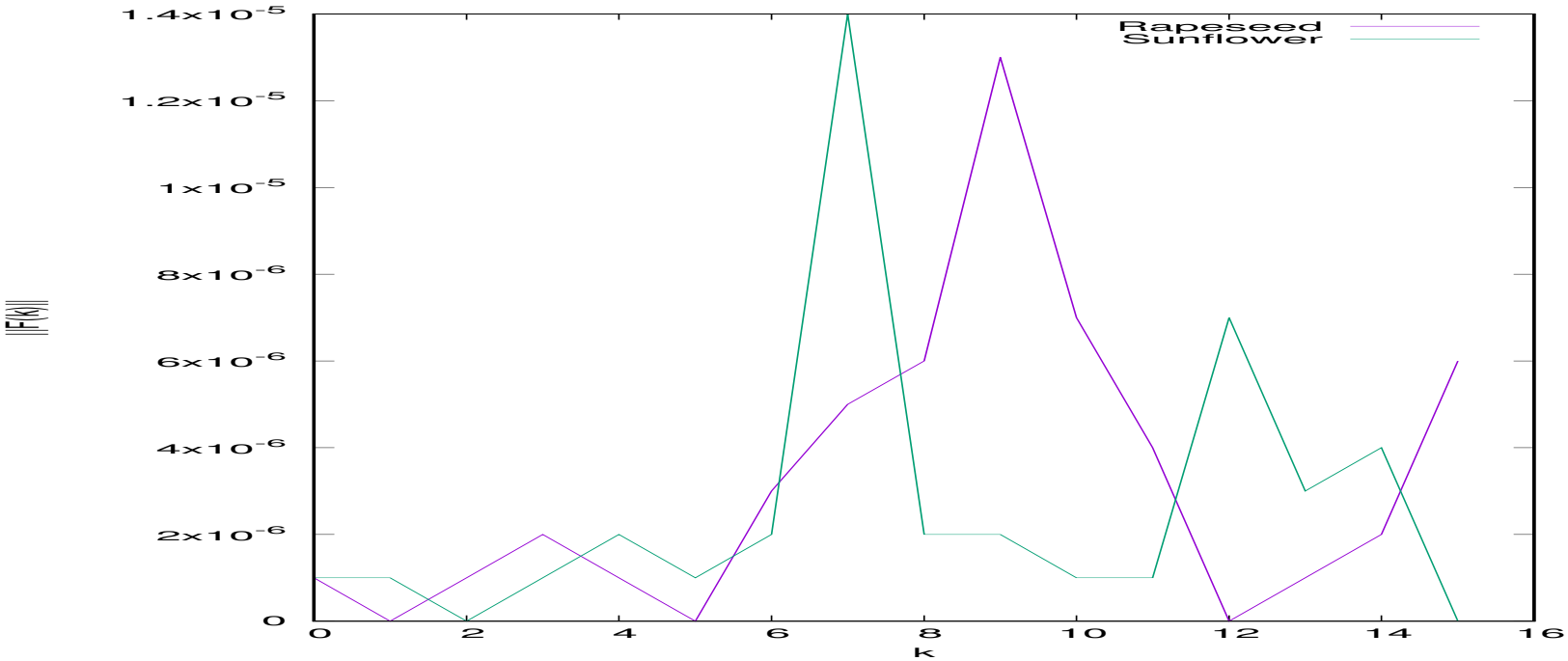
Fast Fourier Transform (FFT) sur l'auto covariance

posons $f_i(\tau) = R_{XX}(\tau)(i,i)$, la FFT discrète $\mathcal{F}_i(k)$ sur les $N = 32$ points est définie par :

$$\mathcal{F}_i(k) = \sum_{\tau=0}^{N-1} f_i(\tau) \exp -j\frac{2\pi k\tau}{N}, \quad k = 0, \cdots N-1 \tag{4}$$

Les valeurs $\frac{k}{N}$ sont appelées les fréquences
$\frac{N}{k}$ représentent les périodes ou temps de retour chez les agronomes

FFT



Spectrogramme vue comme une fonction de k/N sur les signaux appartenant aux items "S" et "R".

Dans une fenêtre de longueur ($N = 32$) chaque item montre une fréquence de $\frac{8}{32}$ soit une période de $\frac{32}{8} = 4$.

Les pics de part et d'autre de $k = 8$ correspondent à la fréquence 1/4.

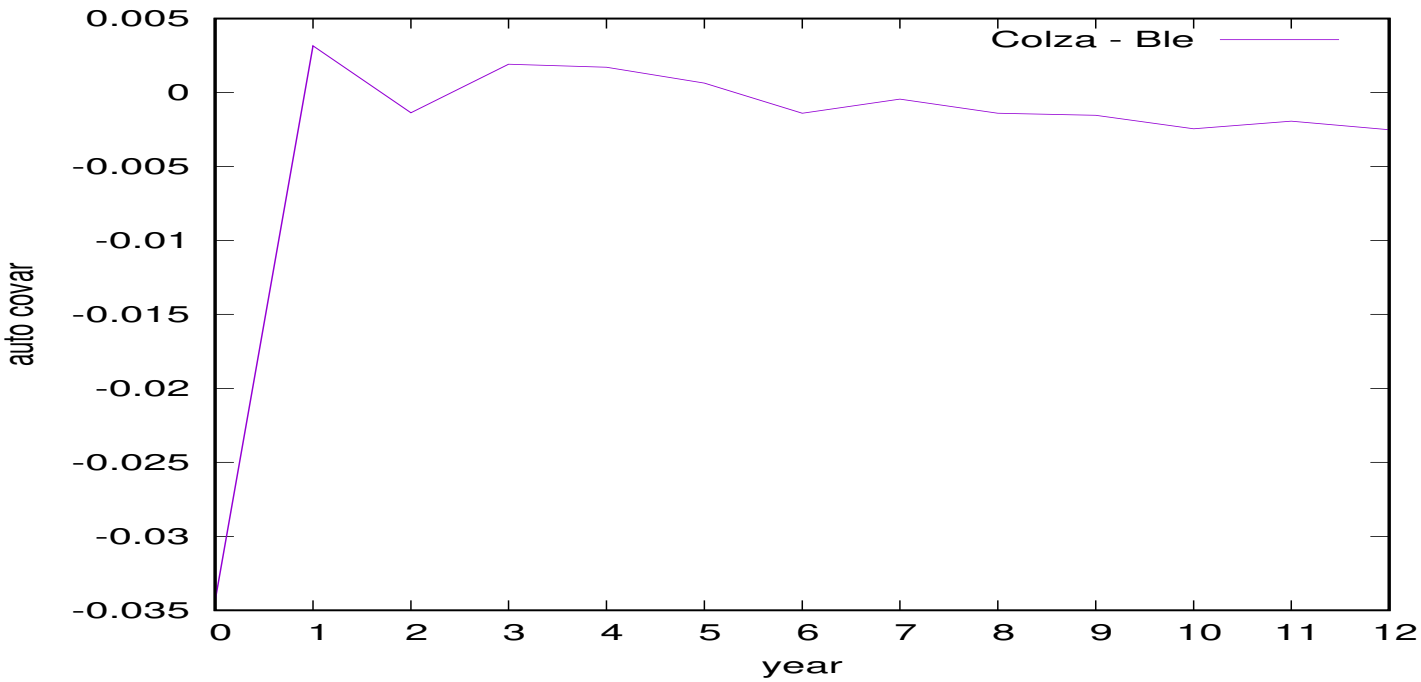
L'imprécision du résultat provient du faible nombre de valeurs $N = 32$ de la série

Conclusion / discussion

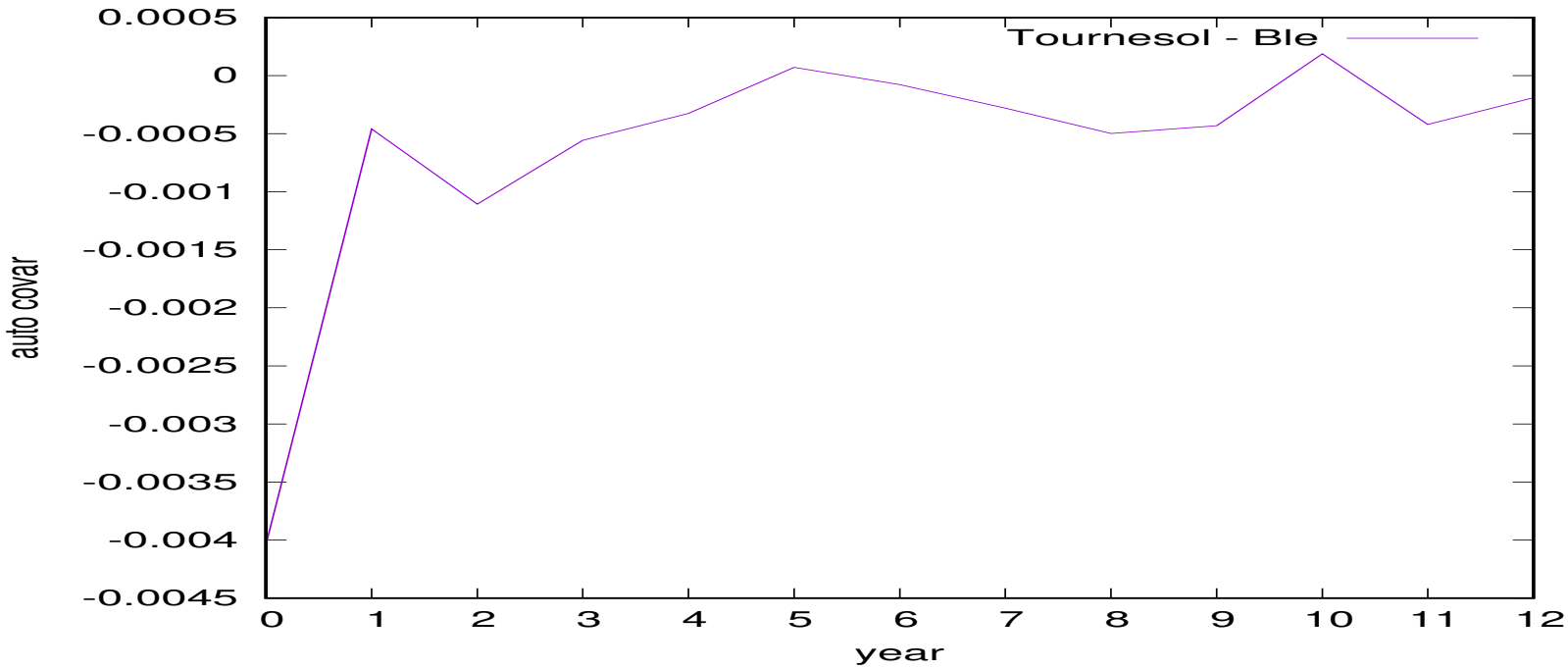
- Notre formalisme autorise l'emploi de la décomposition spectrale pour des données catégorielles ;
- A partir de l'analyse spectrale sur les tenseurs, on fait apparaître des comportements périodiques dans des séries temporelles permettant l'extraction de nouvelles connaissances ;
- L'hypothèse de stationnarité et d'ergodicité est raisonnable tant que l'ensemble des opportunités rencontrées ou contraintes subies par les cultivateurs ne changent pas ;
- On accepte une certaine variabilité inhérente à toute activité humaine ;
- Dans le cadre d'une fouille de données, une interaction serrée avec un agronome permet de détecter et quantifier des rotations entre cultures dans une région agricole.

Application à des signaux réels

Région de Contrexeville - Vittel



Présence de l'itemset Colza - Blé



En plus de l'itemset [Tournesol Blé], une analyse dans les données fait apparaître l'itemset [Tournesol ? ? ? ? Blé] correspondant à la succession [Tournesol (Prairies tempo. x 4) Blé]


M. G. Elfeky, W. G. Aref, and A. K. Elmagarmid.
Periodicity detection in time series databases.
IEEE Transactions on Knowledge and Data Engineering, 17(7):875–887, July 2005.


Esther Galbrun, Peggy Cellier, Nikolaj Tatti, Alexandre Termier, and Bruno Crémilleux.
Mining Periodic Patterns with a MDL Criterion.
In *ECML/PKDD 2018 European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases*, pages 535–551, Dublin, Ireland, September 2018.


Zhenhui Li, Jiawei Han, Bolin Ding, and Roland Kays.
Mining periodic behaviors of object movements for animal and biological sustainability studies.
Data Mining and Knowledge Discovery, 24(2):355–386, Mar 2012.


G. Tatavarty, R. Bhatnagar, and B. Young.

Discovery of temporal dependencies between frequent patterns in multivariate time series.
In *2007 IEEE Symposium on Computational Intelligence and Data Mining*, pages 688–696, March 2007.

Michail Vlachos, Philip Yu, and Vittorio Castelli.
On periodicity detection and structural periodic similarity.
In *Proceedings of the 2005 SIAM International Conference on Data Mining, SDM 2005*, 04 2005.