A strategy to help beekeepers choose breeder colonies using vine copulas

Christophe Reype[1], F. Mondet, T. Opitz, E. Bassi, B. Basso, M. Gotti, G. Guido, G. Kairo, C. Kouchner, P. Jourdan, P. A. Mochet, E. Rottier, M. Tagliabue

[1]INRAE, Abeilles et Environnement, 84914, Avignon, France

christophe.reype@inrae.fr

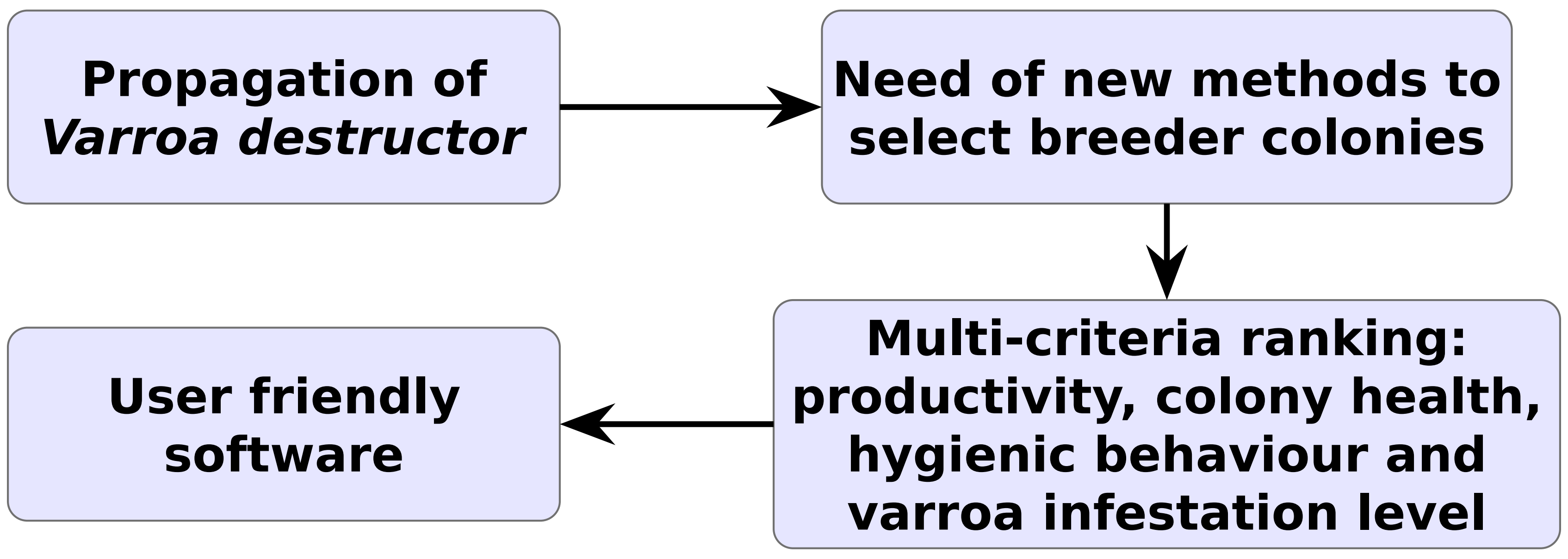
Interreg Cofinancé par l'Union Européenne
France – Italie ALCOTRA
Melior'api

Abstract

Varroa destructor is a parasitic threat to honey bees, often leading to reduced productivity and, if unmanaged, colonies collapse within 2–3 years.

We propose a multi-criteria ranking method to help beekeepers choose colonies for breeding their bee populations. The method combines two components. First, a weighted Euclidean distance enables the comparison of colonies while reflecting beekeeper-defined priorities across multiple criteria, including productivity, colony health, hygienic behaviour, and varroa infestation level. Second, a vine copula approach models the joint distribution of these criteria, capturing their dependencies and allowing for the quantification of uncertainty in the rankings.

Context



Propagation of *Varroa destructor*

Need of new methods to select breeder colonies

Multi-criteria ranking: productivity, colony health, hygienic behaviour and varroa infestation level

User friendly software

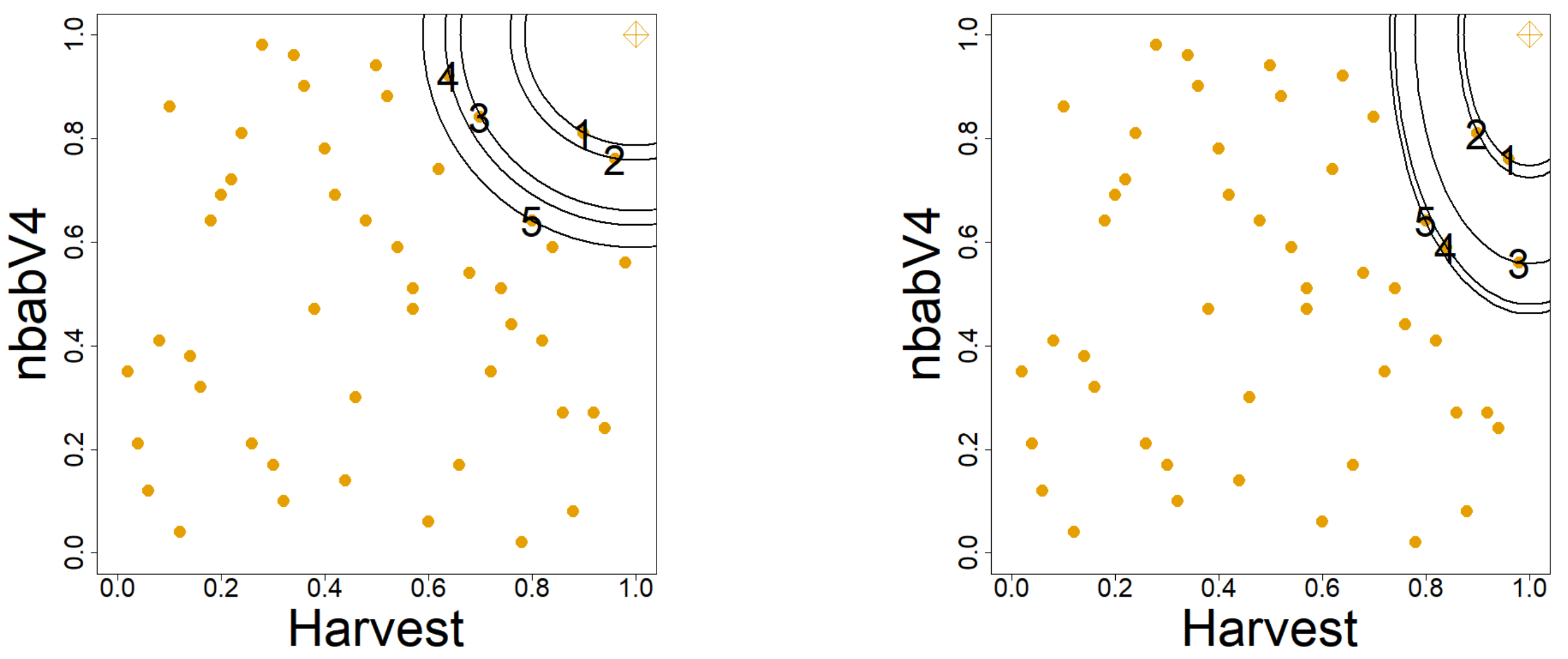
Ranking method

Starting points:

- The perfect hive ($\mathbf{u}^*$) is a hypothetical hive made by the best value of each variable of interest.
- The best hives are the ones that most resemble the perfect hive.
- Each variable is standardised to a uniform scale u_k between 0 (worst) and 1 (best).

Key assumptions:

- The resemblance of the hive i ($\mathbf{u}^{(i)}$) to the perfect hive is based on a weighted Euclidean distance: $d(\mathbf{u}^{(i)}, \mathbf{u}^*) = \sqrt{\sum_{k=1}^{K} w_k (u_k^{(i)} - u_k^*)^2}$ where $w = (w_1, \ldots, w_K) \in \mathbb{R}^K$ are the user-defined weights assigned to each variable x_k and representing their importance.



Top 5 hives of an apiary with the weight $(0.5, 0.5)$ (left) and $(0.8, 0.2)$ (right) according to the normalised variables "Harvest" and "nbabV4" (the number of bees before winter). The dot represent the hives of the apiary and the square is the perfect hive.

Distribution of the data:

- The distribution of the data can be modeled by a copula [2], more specifically a vine copula [1].
- This vine copula is used to stochastically generate a large number ($n = 10^5$) of theoretical hives ($\mathbf{t}_1, \ldots, \mathbf{t}_n$).
- The probability of having a better theoretical hive than the hive i, also called risk for the beekeeper, can be estimated by:

$$r(\mathbf{x}^{(i)}) = \frac{1}{n} \sum_{j=1}^{n} \mathbb{1}\left\{ d(\mathbf{t}_j, \mathbf{u}^*) \leq d(\mathbf{u}^{(i)}, \mathbf{u}^*) \right\}. \qquad (1)$$

- The hives are ranked according to their risk.

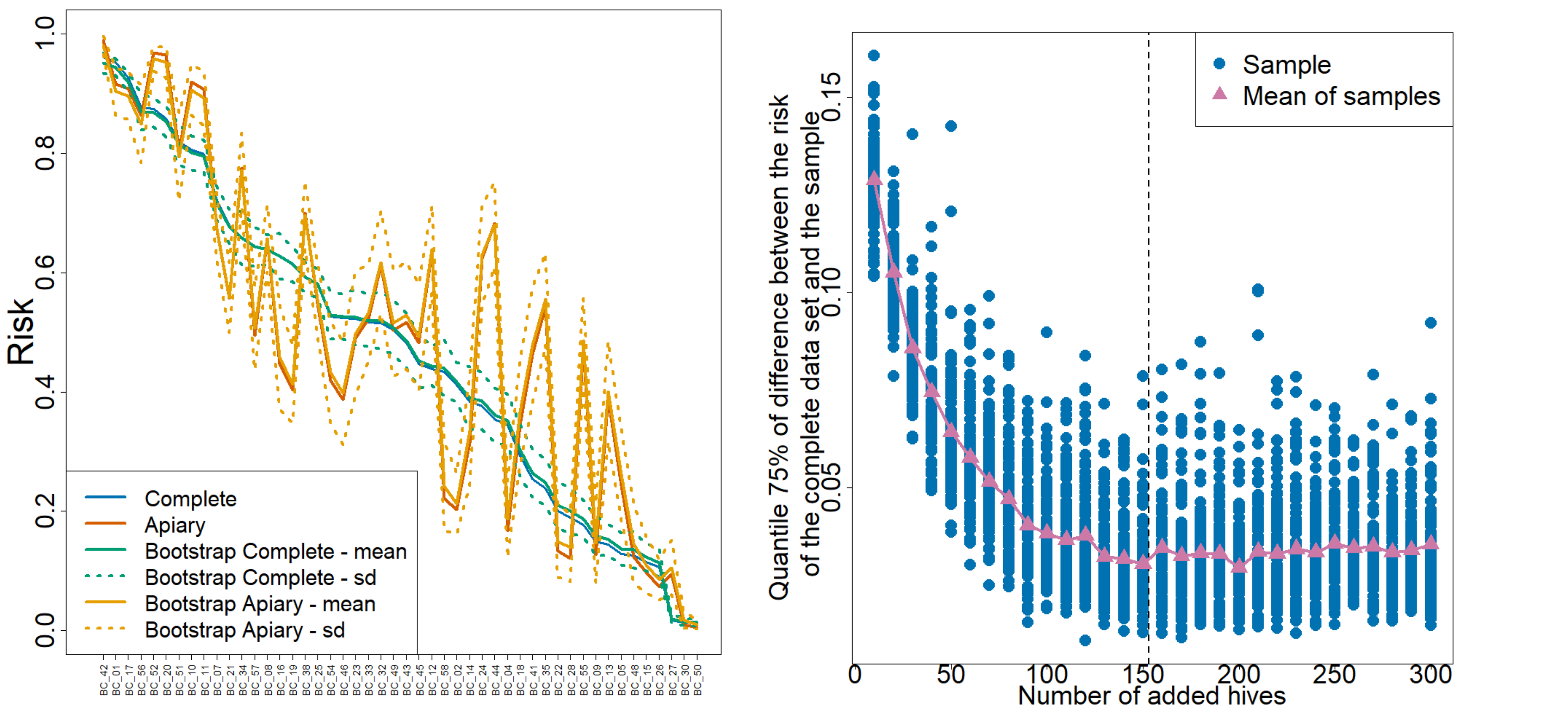
Vine Copula

- According to Sklar's theorem [2], there exists a function $C : [0,1]^K \to [0,1]$, called the copula and its associated density c, such that for $\mathbf{x} = (x_1, \ldots, x_K)$, the cumulative data distribution is:
$$F(\mathbf{x}) = C\big(F_1(x_1), \ldots, F_K(x_K)\big).$$
- The vine copula [1] is a parametric and multivariate model constructed from multiple bivariate copulas.
- Model selection is based on maximum likelihood and AIC.

Results

Application:

- A vine copula is fitted on the data of 1 apiary (orange) and another on the data of 8 apiaries (green).
- The bootstrap samples are used to test the robustness of the estimation (e.g. its mean and standard deviation).
- Bootstrap samples are used to test the effect of the number of hives on the computation of the risk.



Risk computed from the data of the apiary (orange) and the complete data set (green), and the standard deviation computed from corresponding bootstrap samples (left). The risk is computed on samples made of the apiary and a bootstrap sample on the rest of the data. The absolute difference between the risk on the complete data set and these samples are computed. The third quartile of the difference between risk is represented (right).

Interpretation:

- The best hives are detected by both vine copulas.
- The hives have relatively close rank from both data sets.
- Risk stability increases with the number of considered hives.

Conclusion and perspectives

- The weighted distance gives a mathematical framework for the empirical ranking made by beekeepers.
- The method uses an easily interpretable ranking criteria.
- The copula takes into account the correlation between the variables.
- The estimation of the copula requires big data sets. Thus, the need for cooperation between beekeepers.
- An app is in developpement to be of use to the community.

References

1. T. Bedford, R. M. Cooke, *The Annals of statistics* **30**, 1031–1068 (2002).
2. M. Sklar, presented at the Annales de l'ISUP, vol. 8, pp. 229–231.