Summary Team 2

ch.voinot

March 2025

The **Allergen Chip Challenge** project aims to characterize molecular allergen profiles and analyze their correlation with food and respiratory allergies.

1 Context and Objectives

The primary objectives of this study were:

- To model and explain variations in IgE levels based on age, region, severity, and allergy type.
- To determine IgE thresholds associated with specific food allergies.
- To analyze correlations between different IgE markers to better understand allergic profiles.

2 Methodology

The study utilized the following analytical methods:

- **Principal Component Analysis (PCA)**: Used to summarize IgE correlations and identify the most influential factors in the dimensions that best capture data variability.
- **ANOVA (Analysis of Variance)**: Applied to explain the total variance in IgE levels across all variables and to identify significant differences in mean IgE levels.
- **Logistic regression and AIC measurement** (lower AIC indicates a better model fit): Used to determine relevant IgE thresholds for allergy diagnosis.
- **Zero-inflated Gamma Regression**: A modeling approach designed to account for skewed distributions and excess zero values in IgE data.

Due to time constraints, our analysis focused exclusively on Chip1, with Chip2 data not being included in this study.

3 Key Findings

1. **IgE Levels and Age**
   PCA revealed that IgE profiles evolve with age, with notable differences among children (0-10 years), adolescents (10-15 years), adults (15-45 years), and seniors (¿45 years). PCA also identified groups of IgE markers that can be summarized under single representatives (e.g., Ara.h markers for food allergies, and Fel.d and Der.p for airborne allergies).
   ANOVA confirmed these findings, showing significant changes in IgE concentrations in the 10-15 age group (+2.5 on average for Ara.h IgE) and a decline in individuals over 45 (-2.6).

2. **Regional Influence on Allergies**
   PCA also highlighted regional variations in IgE profiles. The South-West and South-East regions showed stronger correlations with pollen allergies (primarily represented in PCA dimension 1), while the North was more associated with fruit allergies (significantly correlated with PCA dimension 2).
   ANOVA on Ara.h IgE, a key marker for food allergies, revealed a significant regional impact. However, the estimates did not confirm the PCA findings for the North but instead highlighted

Bourgogne-Franche-Comté (+26 on average for Ara.h IgE), Centre-Val de Loire (+16), and Outre-Mer (-7) as having notable effects.

3. **Thresholds for Diagnosing Allergies**
   Logistic regression models were applied to determine optimal IgE thresholds for allergy detection, minimizing AIC values. Initially, thresholding results were poor but improved with variable selection. Age was identified as a crucial predictor for a subset of IgEs (particularly in the joint prediction of Gal, Ara, and Bos IgEs).
   These findings align with PCA results, which also demonstrated the impact of age groups on the first two principal components. However, further refinement is needed to establish conclusive IgE thresholds.

4. **Modeling IgE Levels with Regression Approaches**
   Traditional regression techniques proved inadequate due to the highly skewed distribution and excessive zero values in IgE data. A zero-inflated Gamma regression model was introduced as a more suitable approach. This model assumes that IgE levels (e.g., Gal.d.2) are either zero with probability p or follow a Gamma distribution where the rate depends on age.

4 Recommendations

- Extend ANOVA analysis on Ara.h (food allergies) to include IgEs responsible for airborne allergies.
- Improve the IgE threshold model by allowing different thresholds for each IgE and identifying the best threshold combinations.
- Further evaluate the performance of the zero-inflated Gamma regression model.
- Explore **nonparametric methods**, such as Distributional Regression Forests (DRF), to refine predictions and develop techniques to consolidate multiple trees into a single model.
- Extend the work on Chip2