Modeling Type 1 Diabetes Environmental Drivers in Metropolitan France

Business Data Challenge Report

Elea Bordais (ENSAE)

Ismael Dembele (ENSAE)

Paul Toudret (ENSAE)

Patryk Wiśniewski (ENSAE)

Abstract

This study investigated environmental and socioeconomic drivers of regional Type 1 Diabetes (T1D) incidence in metropolitan France using 2023 hospital (PMSI) and aggregate data with regression and machine learning models. Key findings indicate a protective association for sunshine duration (proxying Vitamin D) and a positive association for regional tobacco addiction prevalence (especially in males), alongside socioeconomic factors. While suggesting these factors modulate regional T1D risk, those findings require cautious interpretation and further investigation.

Academic Supervisor: Professor Khaleghi
Business Data Challenge (2025)

ENSAE Paris
Academic Year 2024-2025

Acknowledgements

We would like to sincerely thank all those who contributed to the realization of our Business Data Challenge project on identifying risk and protective factors for Type 1 diabetes.

First of all, we would like to thank the team at Sanofi, particularly Noémie Allali, Olivia Carnapete and Thomas Séjourné, for their continuous engagement throughout the project. Their guidance and the data they provided were essential to the development of our work. Their interest in our progress and their feedback during key stages of the project were particularly valuable.

We would also like to thank our coach, Azadeh Khaleghi, for her insightful advice and for helping us define a clear and rigorous modeling approach. Her perspective and expertise helped us move forward with confidence at each step of the project.

Our gratitude also goes to Michael Visser and Yuanzhe Tang, coordinators of the challenge, for their ongoing support and responsiveness. Their involvement was essential in guiding us through both practical and methodological aspects of the project.

Finally, we would like to thank the Data Storm and CASD teams for facilitating secure and efficient access to the data, which made this work possible.

Table of contents

1 General Introduction

Type 1 diabetes (T1D) is an autoimmune disease characterized by the immune system's destruction of insulin-producing beta cells in the pancreas. Without effective management, T1D can lead to severe complications, including diabetic ketoacidosis and hypoglycemia. In France, the disease represents a significant public health issue, with approximately 4,000 new diagnoses annually and an estimated 180,000 individuals living with T1D between 2018 and 2022. The rising incidence, particularly the 4% average annual increase observed in pediatric populations according to Santé publique France, underscores its growing impact.

The origin of T1D is understood to be multifactorial, involving genetic predisposition interacting with environmental triggers. While numerous environmental and socio-demographic factors have been linked to T1D development, their influence can vary significantly across different regions and populations. In France, the prevalence of T1D exhibits notable regional disparities, suggesting that external factors, such as lifestyle, healthcare access, socioeconomic conditions, pollution, and climate, may modulate disease risk. Consequently, the early identification of at-risk populations and a clearer understanding of specific environmental contributors remain critical public health challenges.

This study focuses on inference rather than prediction, aiming to explain the relationships between environmental, social, and economic factors and the incidence of T1D in France. We seek to assess how exposures like air pollution, variations in healthcare service access, and socio-demographic variables potentially influence the likelihood of developing T1D across different geographical areas.

To achieve this, our analysis integrates individual-level hospital data from the French Programme de Médicalisation des Systèmes d'Information (PMSI) with diverse open-access external datasets which includes some data that was provided by Sanofi.

2 Literature Review

For any observational study, it is important to understand what are the main insights from the existing literature. A review of the literature allows us to identify what variables matter in the context of the study of type 1 diabetes and to identify possible research gaps. This section summarizes key findings across several factors, from air pollution and nutrition to stress and socioeconomic conditions.

2.1 Air pollution

Air pollution is among the most studied environmental factors in recent years regarding the risk of T1D onset, especially in children. ozone (O3), nitrogen oxide (NOx) and particulate matter (PM10 and PM2.5) are known for their effects on inflammation, oxidative stress and immune regulation. Numerous studies support that exposure to these air pollutants during pregnancy or early life could contribute to higher risk of developing T1D. A Canadian study carried out by Elten et al. 2020 examined around 750k children and found that maternal exposure to ozone during first trimester of pregnancy is associated to increased risk of early-onset pediatric diabetes. A study in Israel by Taha-Khalde et al. 2021 supported this result by showing that children whose mothers were exposed to high concentration of O3 (third and fourth quartiles) during pregnancy had higher chances of developing T1D. However, Taha-Khalde also investigated the role of PM10 and PM2.5 exposure but found no clear associations after adjusting for socioeconomic status and meteorological conditions. A study carried out in California by Hathout et al. 2006 also found that children with T1D had higher cumulative exposure to O3 from birth, which reinforces the potential association between long-term ozone exposure and autoimmune diabetes. Finally, in Poland, Michalska et al. 2020 explored regional variation in air pollution and T1D incidence. They found a significant positive correlation between annual average PM10 levels and incidence of T1D among children aged between 0 and 18. However, they found no significant correlation between NO2 levels and incidence of T1D in this population. The literature seems to suggest that NO2 plays a secondary role compared to oxidative pollutants like ozone and fine particulate matter.

2.2 Vitamin D and sun exposure

Another widely researched factor is vitamin D, it is seen as a protective factor because of its role in the immune system and pancreatic function. The active form of vitamin D (1,25(OH)2D3) attaches to receptors (VDR) present immune cells and beta cells, where it helps regulate immune responses by limiting inflammation and promoting immune tolerance. Many observational studies suggest that vitamin D supplementation during childhood might lower the risk of developing T1D. A large Finnish birth cohort study by Hyppnen et al. 2001 found that supplementation of vitamin D during childhood was associated with an 88% lower risk of developing T1D compared to children who didn't receive vitamin D. The EURODIAB case control-study confirmed this result, showing vitamin D supplementation in infancy was associated with 33% lower odds of developing T1D later in life. Vitamin D is primarily produced in the skin through exposure to ultraviolet B (UVB) rays from sunlight, making sun exposure an essential component for maintaining high level of vitamin D. Geographic differences in UVB quantities, determined by latitude and cloud coverage, can therefore influence vitamin D status in population. A studied by Mohr et al. 2008 carried out during several years among children aged 0-14 in 51 regions worldwide found that higher regional UVB irradiance was strongly associated with lower incidence rates of T1D in children. The study observed a clear variation with latitude: regions closer to the equator, with greater UVB exposure, had significantly lower T1D incidence compared to regions farther from the equator. Despite consistent results across numerous studies, these results don't establish a causal relationship between supplementation of vitamin D during childhood and development of T1D. Most studies are observational and rely on retrospective data, often based on parental recall of supplementation, without objective measurements of vitamin D levels. There are still many questions about ideal dose, timing and duration and heterogeneity effects.

2.3 Dietary factors

Dietary factors are known to have an impact on the onset of Type 2 diabetes, but they have also been examined in relation to T1D, especially through its impact on immune development and pancreatic beta-cell stress. Several dietary factors during pregnancy, infancy or childhood could contribute to either increasing or reducing risk of developing Type 1 diabetes. Early exposure to certain foods, such as cow's milk or processed meats, has been associated with a higher risk, possibly due to the presence of inflammatory compounds like nitrites and advanced glycation end-products that may promote beta-cell damage (Virtanen 2016). Many prospective and case-control studies have investigated the effects of diet on the onset of T1D. Early introduction of cow's milk and short breastfeeding duration have both been associated with an increased risk of islet autoimmunity and T1D, particularly when they occur in the first months of life (Virtanen 2016). Furthermore, early introduction of solid foods has also been linked to a higher risk of islet autoimmunity in genetically susceptible children. A randomized trial found that giving infants a hydrolyzed formula instead of a standard cow's milk–based formula in the first 6-8 months of life could delay the emergence of islet cell autoimmunity. According to studies, diet continues to play a role in older children. A Swedish study found that children developing T1D had significantly higher intakes of energy, carbohydrates, and especially sugars compared to other children. High sugar consumption remained a significant risk factor after adjusting for total energy intake, suggesting a specific impact beyond general overnutrition. Moreover, Traversi et al. 2020 conducted a small study on children of Italian origin and migrant families in Italy, which included 40 cases and 56 controls. They observed a modest but significant association between higher total caloric intake and the likelihood of developing T1D. In particular, high intake of protein, fat, or carbohydrates individually was also associated with a slightly increased risk of the disease. Evidence suggests that family dietary habits - especially those influencing early nutrition and energy intake - could play a role in shaping T1D risk.

2.4 Psychological stress

In addition to physical exposures, psychological stress has been proposed as a possible trigger for T1D, particularly when stress occurs during sensitive developmental periods. Two studies have studied this relation in the context of acute, traumatic life events. Zung et al. 2012 reported a significant increase in T1D incidence among Israeli children following the Second Lebanon War, suggesting that important levels of psychological stress may act as a trigger. Virk et al. 2010 found that children whose mothers experienced the death of a close family member during pregnancy - particularly in cases of sudden or traumatic loss, such as the death of a partner or a child - were at greater risk of developing T1D, with the association being especially strong among girls. However, the idea that "post-traumatic" type 1 diabetes exists remain controversial. The literature is inconsistent, some studies report a clear link between traumatic life events and diabetes onset, and other studies, including studies conducted in high-prevalence countries like

Sweden, find no such association (Littorin et al. 2001). A key limitation of the literature on the relation between stress and type 1 diabetes is the narrow focus on extreme events such as wars or losses of loved ones. Few studies have focused on chronic and everyday stressors such as school pressure, neighborhood criminality or family instability and their influence on developing T1D.

2.5 Tobacco use and exposure

Tobacco use and exposure have also raised concerns regarding onset of T1D, due to their known effects on insulin secretion and pancreatic beta-cell function. While smoking is well known for its effects on insulin resistance, evidence also points to its role in impairing insulin secretion. A Japanese cohort study by Morimoto et al. 2013, which followed around 2k men over several years, found that current smokers were nearly twice as likely to develop impaired insulin secretion compared to men who never smoke. The risk increased in a dose-dependent manner with the number of pack-years smoked, suggesting a cumulative effect. Interestingly, smoking was not significantly associated with insulin resistance in this cohort, reinforcing the idea that smoking may directly damage beta cells rather than simply altering insulin sensitivity . A Swedish study confirmed those results showing that smokers had significantly lower beta-cell function that people who had never smoked, indicating reduced insulin secretion capacity. Notably, this association was not observed in women, suggesting possible sex-specific biological responses to tobacco exposure (Ostgren et al. 2000) While these studies show that smoking can impair beta-cell function, they do not clearly link smoking to the development of T1D itself. Most research focuses on adults and does not address early-life exposure, when T1D typically begins.

2.6 Socioeconomic factors

Lastly, socioeconomic factors are increasingly studied as potential non-genetic contributors to T1D. Several studies suggest socioeconomic factors can influence access to healthcare and exposure to environmental factors and behavioral risk factors, which may affect onset of T1D. The study by Traversi et al. 2020 showed that children receiving regular health check-ups were significantly less likely to develop the disease. Further analyses support the idea that regional deprivation is associated with higher T1D incidence. A large study by Buchmann et al. 2023, using registry data from nearly 25,000 children in Germany, found that districts with very high socioeconomic deprivation (as measured by the German Index of Socioeconomic Deprivation, GISD) had a higher incidence of T1D compared to those with very low deprivation. Another study carried in Germany by Du Prel et al. 2007 showed that higher deprivation scores – based on income, education and professional training – were significantly associated with increased T1D incidence across regions. The study also reported a clear linear trend, with incidence rising as deprivation increased. It is common to use income of the household, education of parents to measure socioeconomic status. However, they may act as proxies for more complex underlying mechanisms. These includes differences in health behaviors, nutrition quality, stress exposure, and healthcare access, which may not be fully captured by standard indicators. Low parental education could be associated with reduced health literacy, leading to delayed recognition of early T1D symptoms or poor diabetes management. Income may reflect a family's ability to access healthy food, stable housing and preventive care, which are factors that could affect the likelihood of developing T1D. Despite the associations found in these studies, there is no definitive evidence that low socioeconomic status directly causes T1D. Most available studies are observational and cannot account for individual-level confounding or causal mechanisms. Moreover, results are not always consistent across countries and time periods, possibly due to differences in healthcare systems, social safety nets and population genetics.

3 Data Presentation

3.1 Phase 1: *Map of France*

The data preparation phase of this project was conducted in two stages. Initially, we focused on compiling a comprehensive dataset describing France across various geographical levels. The data were sourced from multiple public repositories, all of which are documented in the accompanying README files.

The dataset was constructed at three distinct geographical levels:

1. **Arrondissements administratifs:** This is the primary unit of analysis, comprising 333 (315 without Corsica or DOM-TOM) subdivisions of French departments. These units are sufficiently large to mitigate issues related to sparsity, particularly given the relative rarity of T1D cases.
2. **Bassins de vie:** Defined by INSEE, these units are designed to capture functional living areas based on access to amenities. With over 1,700 such units in France, they offer a finer spatial granularity and are arguably less arbitrary than administrative boundaries.
3. **Departments:** In certain analyses, even arrondissements are too small. For these cases, we aggregate data at the departmental level to ensure sufficient sample sizes.

The variables included in the dataset span a wide array of domains, including meteorological data, air and water quality, socio-economic indicators, crime rates, housing quality and energy efficiency, access to social housing, and business establishment counts. Detailed documentation of these variables is available in the corresponding README files.

3.1.1 PCA on all variables

Despite efforts to retain only contextually relevant variables during preprocessing, approximately 120 features per geographical level were retained. This high dimensionality posed challenges for preliminary analyses. Therefore, we conducted an exploratory Principal Component Analysis (PCA) to identify the primary axes of variation across our *Map of France*. This approach not only facilitated the descriptive analysis but also informs subsequent variable selection.

We report PCA results for the arrondissement level, as the principal components observed were broadly consistent across both the arrondissement and bassin de vie levels.

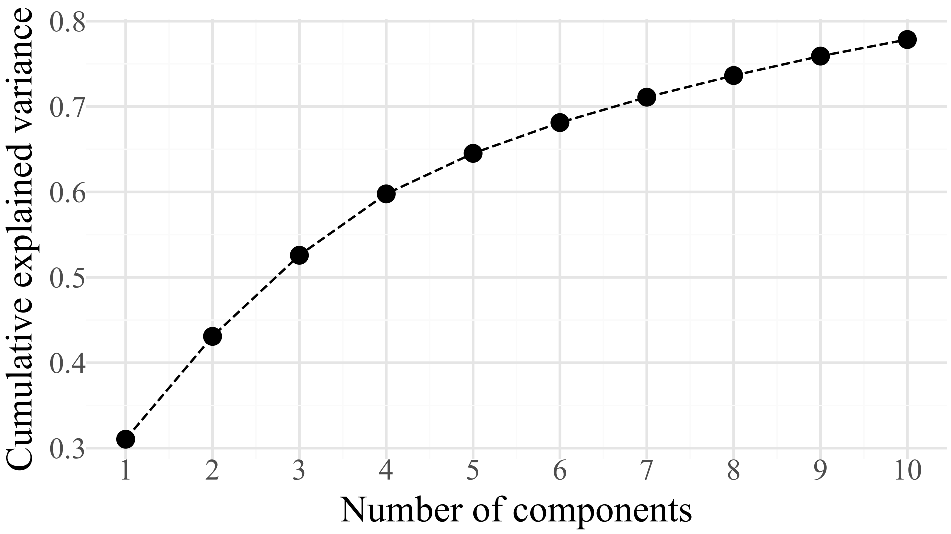

Figure 1: PCA Explained Variance, All Variables

The plot above illustrates the proportion of variance explained by each principal component (PC). For the purposes of descriptive analysis, the results are acceptable: the first four components account for approximately 60% of the total variance. However, even after including ten components, the cumulative explained variance reaches only around 80%, which is insufficient for use in predictive modeling. Therefore, this PCA is employed strictly for exploratory and descriptive purposes.

The main axis of variations we identified with variable loading are the size of an arrondissement (dim. 1), accessibility of services (dim. 2), wealth (dim. 3), housing and weather (dim. 4) and air quality (dim. 5). Some variables are very often highly relevant and as such they should be given a somewhat heavier consideration in the selection process.

The most influential variable loadings are presented along a cartography of France along the predefined axis in appendix A.1.

3.1.2 PCA for social and economical variables

Despite the limitations of global PCA for modeling, we believed that dimension reduction could be more effective when applied to thematic subsets of variables, particularly those related to socio-economic status (SES). Drawing from established practices in sociology, we aimed to condense SES indicators into a small number of interpretable dimensions.

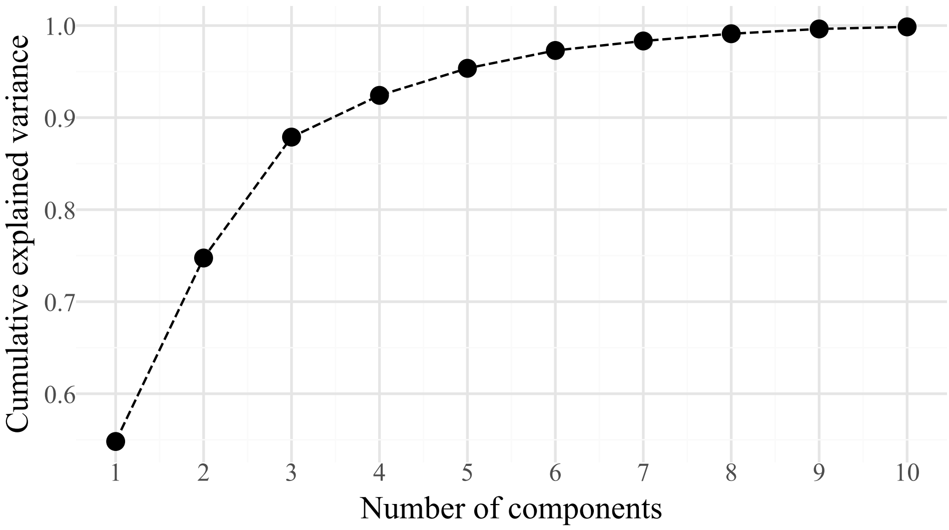

Figure 2: PCA Explained Variance, SES Variables

This approach proved successful: nearly 90% of the variance in SES variables was captured by just three principal components, each of which exhibits clear interpretability:

1. **PC1:** Captures arrondissements characterized by high levels of educational attainment and high income levels.
2. **PC2:** Represents arrondissements with relatively low educational attainment but still moderately high wages.
3. **PC3:** Reflects arrondissements with low educational attainment, with minimal contribution from income-related variables.

Given the interpretability of these components, they will be retained for inclusion in our econometric models. The same exercise was conducted at the bassins de vie level, and it proved equally successful.

1. **PC1:** Captures BV characterized by high levels of educational attainment and high income levels.
2. **PC2:** Represents BV with relatively low educational but with a high first decile and median income.
3. **PC3:** Reflects BV with average educational attainment and low income inequalities.

3.2 Phase 2: Health Data

Through access to PMSI health records via the *CASD* secure platform, we were able to identify patients affected by T1D using both principal and associated diagnosis codes ranging from E100 to E109. This subset constitutes what we refer to as the prevalent population, comprising over 100,000 individuals who had at least one hospital visit associated with a T1D diagnosis. While informative, prevalent populations offer limited insight into risk factors due to several confounding dynamics. Patients may relocate across geographical units over time, and, crucially, the behavior and environment of diagnosed individuals often differ significantly from those who are undiagnosed, introducing concerns of reverse causality in modeling efforts.

Although the datasets are rich in clinical detail, they are limited temporally to 2023. This restricts our ability to robustly estimate incidence rates based on newly diagnosed cases. Given this constraint, we aimed to approximate the incident population by focusing on a more specific diagnostic subgroup.

We selected individuals coded with E101, corresponding to episodes of diabetic ketoacidosis, a serious and typically initial presentation of untreated T1D. This group provides a plausible proxy for incident cases, as such episodes often reflect undiagnosed or newly diagnosed individuals requiring immediate medical intervention. This criterion yielded a cohort of just over 9,000 individuals. We plot the comparative gender and age distribution of both proxy *incident* and *prevalent* populations:

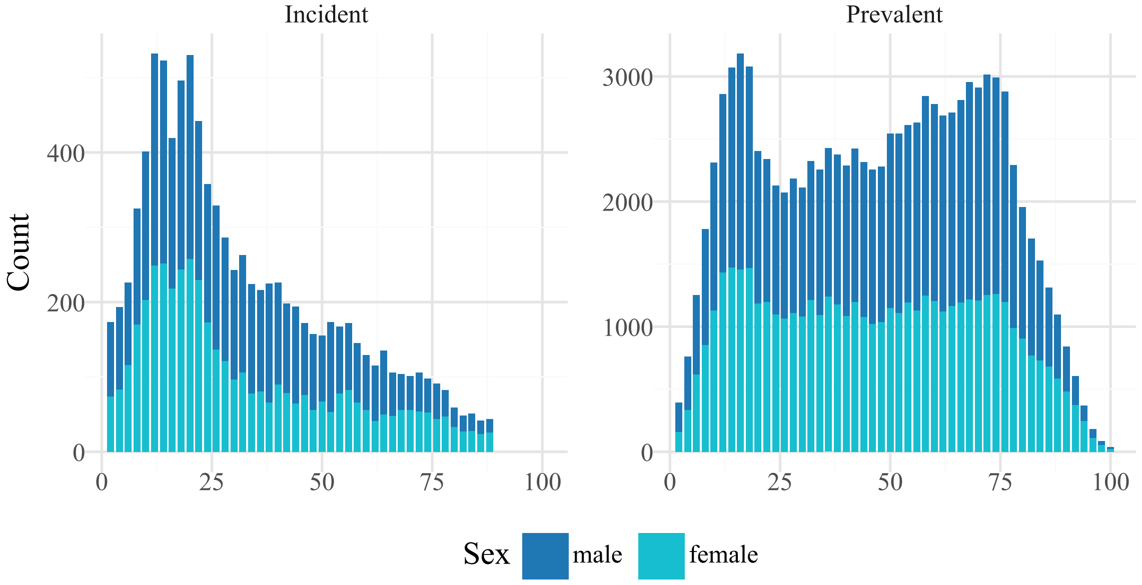

Figure 3: T1D Population, Age and Sex, France 2023

This E101-defined population exhibits markedly different characteristics from the broader prevalent group, most notably in terms of age. Given that T1D predominantly affects younger individuals, we further restricted our sample to those aged 29 years or younger. [1] This final subset, our working proxy for the incident T1D population, includes slightly more than 5,000 individuals. We subsequently compared the temporal distribution of hospital visits across the defined populations and observed marked differences in their admission patterns.

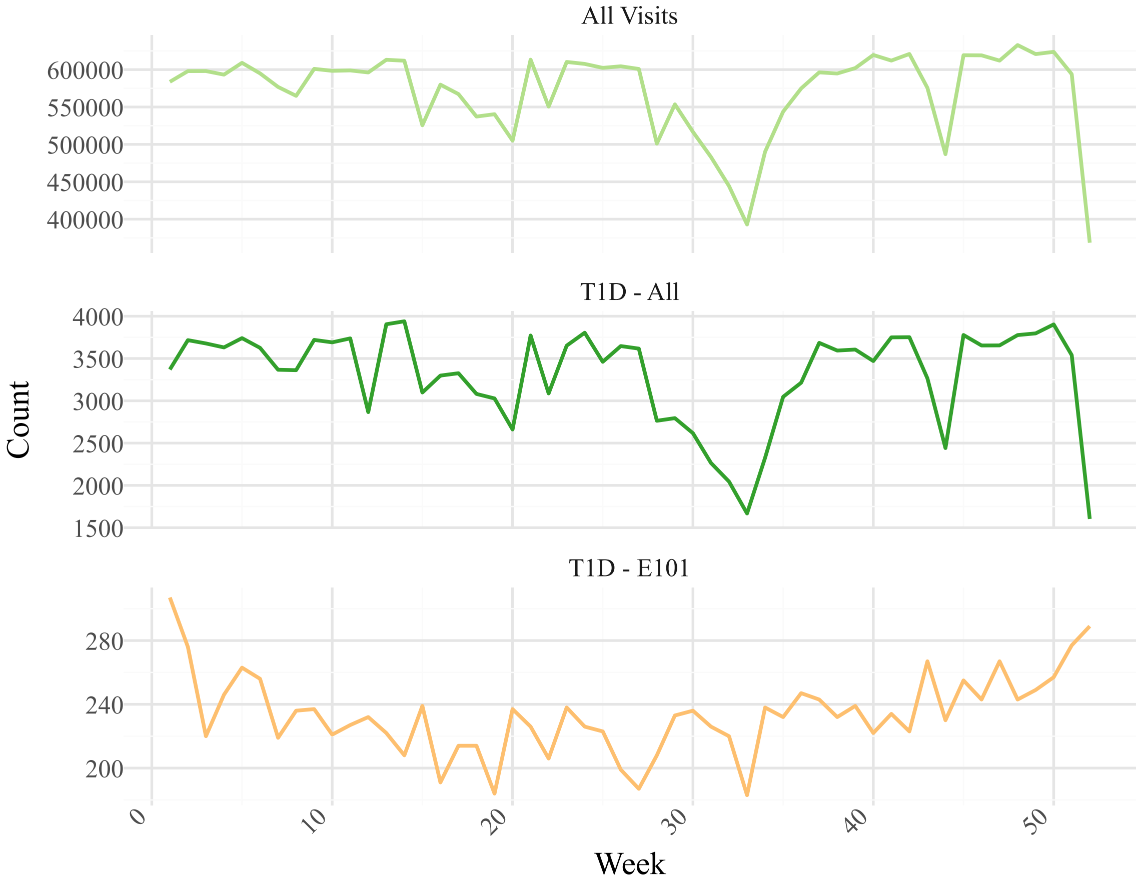

Figure 4: Distribution of visits, T1D population France, 2023

Notably, the prevalent population exhibits a temporal distribution of hospital admissions closely aligned with that of the overall hospital population. In contrast, the incident population displays a pronounced U-shaped distribution: admission rates are lowest during the summer months and peak during the winter. Finally we wanted to see if some associated diagnosis were more likely to appear in our different populations, and again we observe significant heterogeneity. We report a few the relationship we found most interesting.

[1] A more conservative age threshold (e.g., 14 years and under) is available through INSEE data; however, we would miss the end of the peak of people admitted in a ketoacidosis state seen in Figure 3.

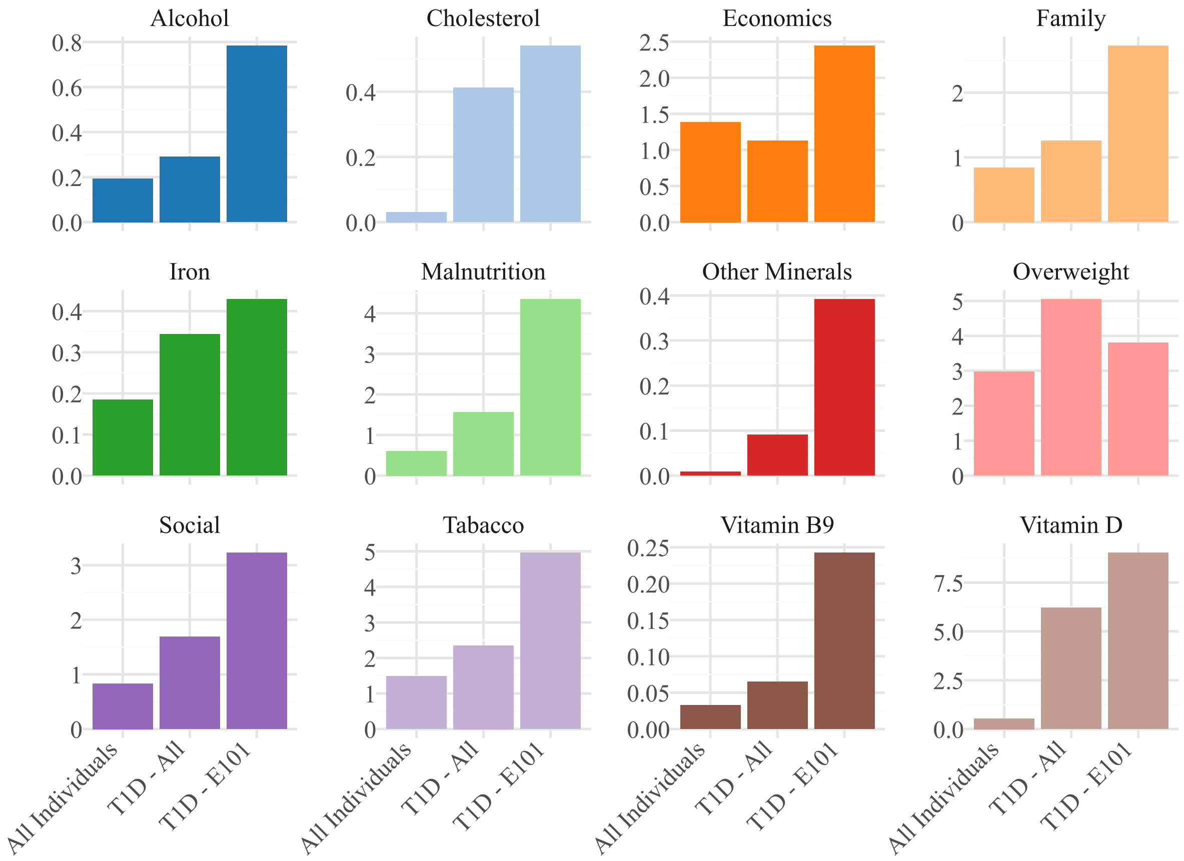

Figure 5: Frequency of associated diagnosis given the population

Individuals admitted in a state of diabetic ketoacidosis were significantly more likely to receive co-diagnoses related to social, economic, or familial difficulties compared to the broader hospital population. These patients also showed higher rates of malnutrition, though not of obesity—unlike the prevalent T1D population, in which overweight and obesity are more common, likely reflecting weight gain associated with insulin therapy (Russell-Jones & Khan).

We also observed a notably high prevalence of Vitamin D deficiency in the incident population. This finding may be linked to the seasonal pattern of hospital admissions discussed earlier, suggesting a potential environmental component to disease onset. Other nutritional deficiencies were identified but are not visualized here due to data sensitivity and re-identification concerns.

Additionally, we found a higher incidence of substance use disorders, particularly related to tobacco and alcohol. Further checks revealed that alcohol-related co-diagnoses were concentrated among older individuals, suggesting these cases may reflect poor disease management—where alcohol use potentially exacerbates or signals lapses in treatment adherence resulting in ketoacidosis. In contrast, tobacco use was more uniformly distributed across age groups.

4 Methodological Overview

Selecting variables for the primary model presented a significant challenge due to the numerous features available in the rich dataset created during this project. While automatic variable selection techniques, such as the elastic net, were considered, they are ill-suited to this research's primary goal: inference rather than prediction (as established in the introduction).

Employing variable selection methods driven by relationships within the dataset itself compromises subsequent statistical inference. Most notably, it renders p-values unreliable for controlling the risk of Type I errors. Furthermore, automatic selection introduces interpretative difficulties. In observational settings often characterized by unobserved heterogeneity, it becomes challenging to discern the substantive reasons for a variable's inclusion beyond its predictive performance or simple correlation with the outcome (here, T1D).

Therefore, we adopted a manual selection approach. This allows us to incorporate variables based on pre-defined hypotheses regarding their potential influence, informed by theoretical considerations from the literature, results from our exploratory PCA (which excluded T1D-related variables), and descriptive statistics concerning seasonality and associated diagnoses. For transparency, the justification for each chosen variable is provided in the annex to this report (appendix A.3). We add or remove certain controls after modeling to evaluate sensitivity of our estimates.

Our primary analysis employs Poisson and Negative Binomial regression models to explain the number of "new cases" observed at various geographical levels. To address inherent concerns about the ecological fallacy, where aggregate-level associations may not accurately reflect individual-level risks, we leverage data across different geographic levels and conduct heterogeneity analyses within selected sub-populations. Finding consistent results across these different levels and groups strengthens the evidence for the robustness of inference.

However, given the limitations common to observational studies, our initial findings are still sensitive to the specific set of control variables included in the model. Therefore, we identify promising explanatory variables from these primary models to investigate further using methods designed for more robust causal inference, treating these variables conceptually as "treatments."

Specifically, we conduct subsequent analyses using either:

1. **Panel data models with fixed effects**: This approach utilizes longitudinal data to control for time-invariant unobserved characteristics of the geographical units.
2. **Double/Debiased Machine Learning (DDML)**: This technique uses machine learning to flexibly control for a potentially large number of confounding variables, aiming to provide less biased estimates of the "treatment" effect. While mechanically different, DDML shares a fundamental goal with methods like propensity score matching: estimating treatment effects by creating comparable conditions between units exposed and unexposed to the "treatment." Furthermore, DDML inherently aids in managing high-dimensional controls through a predictive first-stage estimation.

5 Results and Comments

As previously mentioned, we employ the widely used Poisson regression model to analyze the number of "incident" cases within specific French geographic units, the *bassins de vie 2022* (BV) and *arrondissements administratifs*. This model assumes that the case count Y_i in unit i, conditional on a set of covariates X_i, follows a Poisson distribution: $Y_i|X_i \sim \mathcal{P}(\lambda_i)$. Model parameters are estimated using maximum likelihood estimation. The core of the model links the expected number of cases to covariates and population size through the conditional mean:

$$E[Y_i|\mathbf{X}_i, \text{pop}_i] = \lambda_i = \text{pop}_i \cdot \exp(\mathbf{X}_i^T \boldsymbol{\beta})$$

Here, pop_i denotes the reference population under 30 years of age within the geographic unit i, serving as an offset that scales incidence risk to the relevant population size. The vector $\mathbf{X}_i$ includes both the control variables and variables of interest defined previously.

A key assumption of the Poisson model is the equality of its conditional mean and variance: $E[Y_i \mid \mathbf{X}_i] = \text{Var}(Y_i \mid \mathbf{X}_i) = \lambda_i$. However, our data exhibit evidence of a slight overdispersion (variance exceeding the mean), with calculated dispersion indices ranging from 1.2 to 1.5. To account for this, we also estimate a Negative Binomial (NB) regression model. The NB model typically retains the same structure for the conditional mean μ_i as the Poisson model but incorporates an additional dispersion parameter α to allow for greater variance:

$$\mu_i = \text{pop}_i \cdot \exp(\mathbf{X}_i \boldsymbol{\beta}) \quad \text{Var}(Y_i|\mathbf{X}_i) = \mu_i + \alpha\mu_i^2$$

This approach offers greater flexibility in modeling the variance compared to the standard Poisson model. While estimating the NB model can sometimes lead to instability, particularly with smaller sample sizes, it directly addresses the issue of overdispersion. For comparison, estimates from both the Poisson and NB models are presented in the following table, where (1) is the NB model at the arrondissement level, (2) is the NB model at the BV level, (3) is a Poisson model at the arrondissement level and finally, (4) is a Poisson model at the BV level.

5.1 Main Poisson and NB

Table 1: Main Regression Results

	(1)	(2)	(3)	(4)
Socio-economic PC1	1.0011	0.96434***	1.00276	0.96524***
	0.9262	0.0015	0.7813	0.0003
Socio-economic PC2	0.99695	0.9514**	1.00218	0.96168*
	0.8868	0.0337	0.9051	0.0579
Socio-economic PC3	1.06621***	1.00051	1.07187***	0.99136
	0.0013	0.9804	0.0	0.6273
Share of Homes with AC	2.69198*	1.55551	2.19283	1.2701
	0.0978	0.1855	0.1273	0.4143
Assaults (per 1k)	1.05691**	1.04763	1.06043***	1.0487
	0.0167	0.2587	0.0004	0.2093
Share w/ Calcium Def.	1.08415	1.37205**	1.09553	1.35973***
	0.4398	0.0155	0.314	0.0078
Share w/ Tobacco Addic.	1.07036***	1.03436*	1.07237***	1.03409**
	0.0077	0.0724	0.002	0.0475
Share w/ Vitamin D Def.	1.09732	1.03616	1.11687*	1.03458
	0.1988	0.4053	0.0681	0.3697
Butchers (per 1k)	1.84047**	1.05507	1.73818**	1.09682
	0.0349	0.7435	0.0333	0.5452
Fishmongers (per 1k)	0.93924	0.80079	0.95952	0.72908

Continued on next page

	(1)	(2)	(3)	(4)
	0.8798	0.3495	0.9086	0.1584
Fast Foods (per 1k)	1.02338	1.02698	1.00688	1.04175
	0.6495	0.3977	0.8706	0.1584
Gyms (per 1k)	0.94446	0.94531*	0.94451	0.93976**
	0.1916	0.0542	0.1369	0.0183
NO2 Concentration (air)	0.9781*	0.99084	0.97804**	0.98052**
	0.0891	0.4405	0.0427	0.0285
O3 Concentration (air)	1.00416	0.99824	1.00349	0.99595
	0.6952	0.8184	0.7006	0.51
PM10 Concentration (air)	1.0153	1.00138	1.02227	1.01882
	0.5584	0.9439	0.3098	0.2248
AIC	1745.325	4649.565	1753.395	4670.682
DF	282.0	1602.0	282.0	1602.0
Dispersion	-	-	1.513	1.197
Pseudo R^2	0.301	0.314	0.096	0.023

The regression coefficients are reported as Incidence Rate Ratios (IRR), calculated $\exp(\beta_x)$. An IRR greater than 1 indicates a positive association with the incidence rate, while an IRR less than 1 indicates a negative association. Given the limited sample size, we consider three significance levels: 0.1, 0.05, and 0.01, denoted by one to three stars, respectively. Full regression outputs are available in Appendix A.5.

5.1.1 Socioeconomic Principal Components

At the arrondissement level, PC3, which we interpret as capturing predominantly low to middle education, is statistically significant and associated with an increase in incidence rates. At the BV level, PC1, representing high education and high income, is significant and negatively associated with incidence. Also at the BV level, PC, defined as middle education with high first decile income, is significant (but less than PC1) and negatively associated with incidence. These support our hypothesis that the correlations between income and incidence observed in prior research likely reflect strongly consumption habits (which are closely tied to education), rather than living conditions per se, which are more directly influenced by income alone.

5.1.2 Health Indicators and Lifestyle Factors

A few variables consistently stand out. The percentage of the hospital population diagnosed with tobacco addiction is significant and positively associated with incidence rates across both BV and arrondissement levels, and in both Poisson and Negative Binomial models.

The number of gyms per 1,000 inhabitants is significant at the BV level. While not statistically significant at the arrondissement level, the similarity in coefficient magnitude suggests this is more likely a matter of limited statistical power than a true absence of effect.

Vitamin D and calcium deficiencies exhibit alternating significance across arrondissement and BV levels. Removing one of the two tends to decrease the p-value of the other. However, it is important to note that diagnostic hospital data on nutrient deficiencies may not be reliable a proxy for population-level nutritional status, as they may also capture prevalent local conditions. To better evaluate the vitamin D hypothesis, we will rely on non-diagnostic variables in a subsequent fixed effects model.

5.1.3 Environmental and Contextual Variables

Butcheries are significant only at the arrondissement level. Interestingly, when this variable is excluded from the model, the number of fast food establishments becomes significant, suggesting some overlapping or substitutive influence on dietary exposure. Assault and battery rates are also significant, but only at the arrondissement level.

NO_2 concentrations exhibit a small but significant negative effect on incidence. This finding is counterintuitive and appears sensitive to the inclusion of control variables. Therefore, we refrain from drawing strong conclusions from this coefficient.

5.1.4 Water Quality and Climate

Variables relating to water quality are not significant with the inclusion of an interaction term with tap water consumption. The proportion of homes with air conditioning (AC) is not always independently significant. However, when interacted with average summer temperatures, AC appears to moderate the positive effect of higher temperatures on incidence. Unfortunately, we are unable to test this interaction at the BV2022 level due to the lack of reliable weather data.

5.2 Poisson Heterogeneity by Sex

To complement the previous estimates and enable comparisons with existing literature, we analyze the heterogeneity of effects by sex. This analysis is based on a restricted sample consisting only of incident patients. As a result, the estimates may be noisier, particularly because some geographic units report zero incident cases. For this reason, we limit the heterogeneous analysis to the arrondissement level. Specifically, Model (1) is a Negative Binomial (NB) regression for females, Model (2) is an NB model for males, Model (3) is a Poisson model for females, and Model (4) is a Poisson model for males.

Table 2: Subgroup Heterogeneity

	(1)	(2)	(3)	(4)
Socio-economic PC1	1.0054	0.99964	1.00612	1.00003
Socio-economic PC2	1.03216	0.97308	1.03083	0.97562
Socio-economic PC3	1.0508*	1.08456***	1.05994**	1.08346***
Share of Homes with AC	3.24361	1.77493	2.78827	1.70811
Assaults (per 1k)	1.06543**	1.05672**	1.0649***	1.05804**
Share w/ Calcium Def.	1.18072	1.02658	1.1696	1.03212
Share w/ Tobacco Addic.	1.03824	1.09889***	1.04239	1.09885***
Share w/ Vitamin D Def.	1.10711	1.10025	1.12871	1.10049
Butchers (per 1k)	1.47949	2.0611**	1.48223	2.00719*
Fishmongers (per 1k)	0.88714	1.04888	0.90995	1.04363
Fast Foods (per 1k)	1.08508	0.95097	1.06569	0.95205
Gyms (per 1k)	0.92865	0.96086	0.92835	0.96166
NO2 Concentration (air)	0.9697*	0.98502	0.97144*	0.98453
O3 Concentration (air)	0.99954	1.00671	1.0005	1.00604
PM10 Concentration (air)	1.01362	1.0255	1.01549	1.02707
AIC	1443.024	1473.68	1443.791	1471.988
DF	282.0	282.0	282.0	282.0
Dispersion	-	-	1.284	1.3
Pseudo R^2	0.298	0.299	0.081	0.063

Consistent with prior literature, tobacco addiction is significantly associated with increased incidence only among males. The effect is not statistically significant in the female-only sample.

A similar pattern of heterogeneity is observed with the presence of butchers, which appears to affect men more than women. One possible explanation is that males may drive demand for butchers to a greater extent than females, though this remains speculative.

The principal component PC3, previously characterized as reflecting lower to middle education, is also more strongly associated with increased incidence among males. This may suggest lower-quality consumption habits among less-educated men.

Finally, the number of assaults per 1,000 inhabitants is positively associated with incidence for both sexes, with a slightly stronger effect observed among females. This may reflect the idea that the feelings of insecurity are more prevalent among women than men.

5.3 Poisson Fixed Effects

5.3.1 Motivation

As previously mentioned, drawing inference from observational data is inherently challenging, especially in the context of ecological inference. Our estimates can vary substantially depending on the inclusion and specification of control variables. For some covariates, we can mitigate this issue by leveraging a time dimension through the use of fixed effects. Fixed effects models are a powerful tool in causal inference, as they absorb all time-invariant unobserved heterogeneity across observational units.

Given the limitations of our data, which includes only one year of observations, only a few variables are suitable for this approach, most notably, monthly weather data. We deliberately avoid using diagnostic hospital data in fixed effects models, as these are subject to seasonal fluctuations in healthcare utilization, potentially introducing population bias. In contrast, weather data should be exogenous and not influenced by such factors. To ensure adequate variability and avoid sparsity (i.e., a high frequency of zeros), we conduct this analysis at the *department* level. Fixed effects estimates are computed using an IRLS-based algorithm with fixed-point acceleration.

The key objective of this FE analysis is to test the vitamin D hypothesis regarding T1D incidence. Since sunlight exposure is crucial for human vitamin D synthesis [Norman 2008], and vitamin D levels typically decline in winter due to limited sun exposure and storage capacity (Holick 2007), we investigate whether increased sunshine is associated with reduced T1D incidence, controlling for other meteorological factors.

5.3.2 Sunshine Effects

Our findings support this hypothesis. Controlling for department fixed effects and other monthly weather variables, the number of sunny days per month shows a statistically significant negative association with the T1D incidence rate (IRR = 0.962, $p < 0.01$). Ceteris paribus, an additional sunny day per month is associated with approximately a 3.8% decrease in the monthly T1D incidence rate within a department. We also tested the inclusion of one-month lags for the weather variables. While the effects appear stronger, they are slightly noisier, due to our short time series. Each additional lag reducing the number of usable observations per department by one.

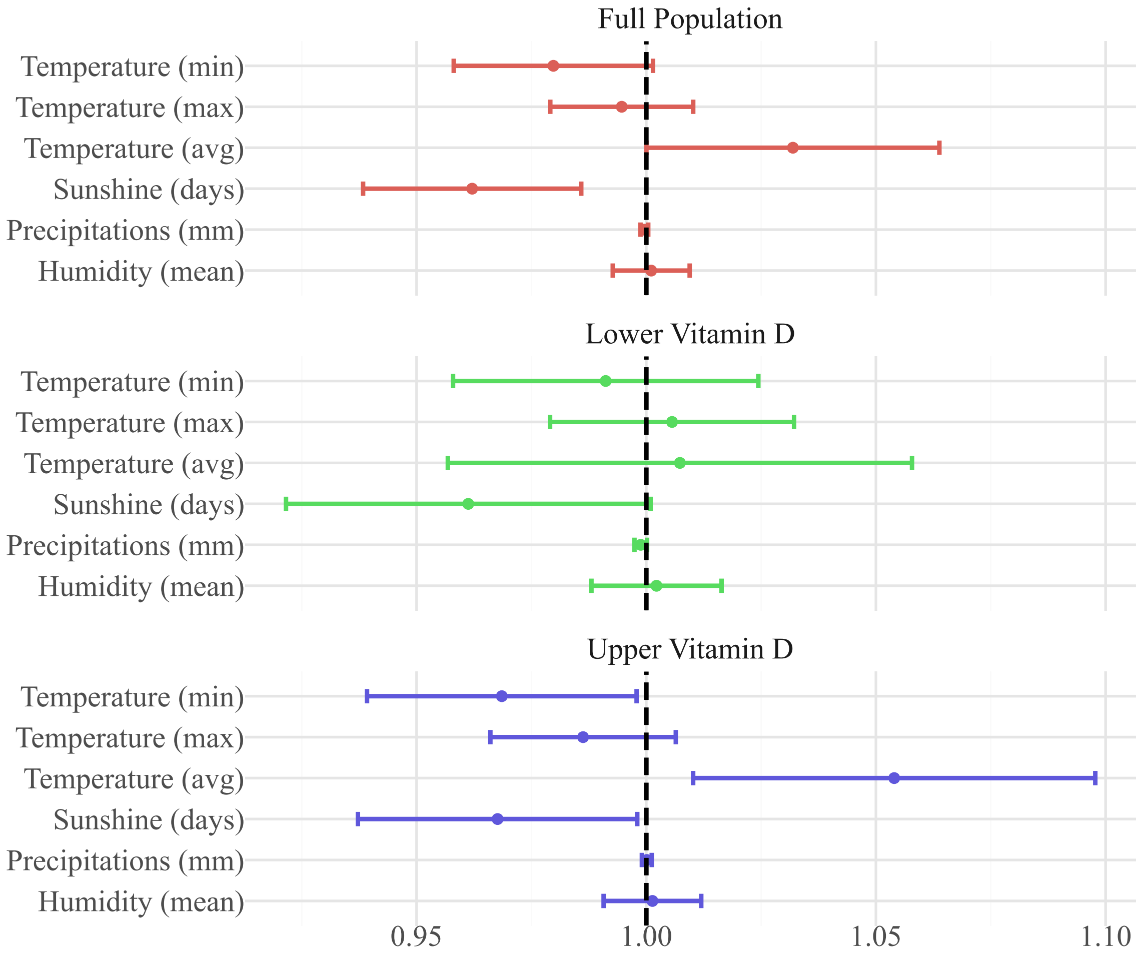

Figure 6: Poisson Fixed-Effects, Consumption Heterogeneity, Clustered SE

5.3.3 Sunshine Heterogeneity

To further strengthen our interpretation, we assess whether the effect of sunshine is more pronounced in regions with lower dietary intake of vitamin D. Such a pattern would be consistent with the idea that sunlight captures vitamin D in our population. Our results indicate that in low-consumption regions, the negative effect of sunlight on T1D incidence is slightly stronger, though also noisier.[2] Nevertheless, the effect remains statistically significant in both high- and low-consumption regions, lending support to the vitamin D interpretation.

Model dispersion is limited, indicating that a Poisson specification is appropriate. The model's pseudo-R^2 ranges from 33% to nearly 50%, largely due to the inclusion of fixed effects. Detailed regression tables can be found in appendix A.7.

[2] Note that the difference is not *statistically* significant

5.3.4 Sunshine and Vitamin D

As a final validation step, we test whether the noisy vitamin D deficiency variable from Section 5.1 is indeed related to sunlight exposure and thus serves as a proxy for population-level deficiency. We estimate a fixed effects linear model:

$$Y_{it} = \alpha_i + \text{sunshine}_{it} + \text{visits}_{it} + \varepsilon_{it}$$

Where the dependent variable Y_{it} is the share of hospital patients diagnosed with vitamin D deficiency in department i at month t. We explain this Y_{it} by a department fixed effect α_i, the number of sunny days and where we attempt to control for seasonality using the number of monthly hospital visits in the department.

Table 3: Sunshine and Vitamin D

	(1)
N. Hospital Visits	-0.00484***
Sunshine (days)	-0.01499***
N. Observations	1128.0
Pseudo R^2	0.925

We find that ceteris paribus, an additional 24 hours of sunshine per month in a department is associated with a 0.015 percentage point *decrease* in vitamin D deficiency, in relative term equivalent to a 2.7% relative decrease. The model's R^2 is 92.5%, with roughly 10% of the variation explained by time-varying covariates. This supports the interpretation that hospital-diagnosed vitamin D deficiency plausibly reflects true population-level variation.

5.3.5 Remark

While these fixed effects analyses significantly strengthen the evidence by controlling for time-invariant confounders, we maintain caution regarding causal claims based solely on this observational study. Establishing a definitive causal link between vitamin D (proxied by sunshine) and T1D would require further research, ideally with longer time series allowing for more extensive time-varying controls. Furthermore, these results represent ecological associations at the department level, which do not automatically translate to individual-level causal effects.

5.4 Double Debiased ML

5.4.1 Motivation

While fixed effects models allow leveraging the time dimension for certain variables, they cannot be applied to time-invariant characteristics or when panel data is unusable for specific factors of interest, such as tobacco addiction patterns derived from hospital data. To estimate the potential causal effect of such variables, we turn to Double Machine Learning (DML). Conceptually, DML, like matching methods, aims to compare observational units that are very similar across a range of characteristics (X) but differ in their level of a specific "treatment" variable (D). We employ DML within a partially linear framework, assuming the outcome (Y, T1D incidence) and treatment (D) models can be represented as:

$$Y = D\theta_0 + g_0(X) + \zeta, \quad \text{where } E[\zeta|D, X] = 0$$

$$D = m_0(X) + V, \quad \text{where } E[V|X] = 0$$

Where, Y is the outcome variable, D is the treatment variable, θ_0 is the target causal parameter, X are the confounding variables, $g_0(X) = E[Y|X]$ and $m_0(X) = E[D|X]$ are unknown nuisance functions, ζ and V are error terms.

A key advantage of DML is its use of machine learning (ML) techniques to flexibly estimate nuisance functions without strong parametric assumptions, which also aids in handling high-dimensional X (variable selection). We utilized Random Forests as the ML learner for both, chosen for its flexibility in capturing complex relationships and its implicit variable selection. This data-driven estimation of nuisance components avoids injecting strong prior beliefs into that part of the model. Further details on the DML methodology are provided in appendix A.4.

This DML analysis was applied to several variables identified as potentially important from the previous models. Due to computational constraints these models were estimated only at the arrondissement level.[3] Checks related to the common support assumption and the predictive performance of the first-stage models are discussed in appendix A.6 and were found to be broadly acceptable.

[3] The analysis was performed on a virtual machine with limited resources (2 vCPUs, no GPU), preventing parallelization of cross-validation or extensive hyperparameter tuning for the Random Forest learners.

5.4.2 Results

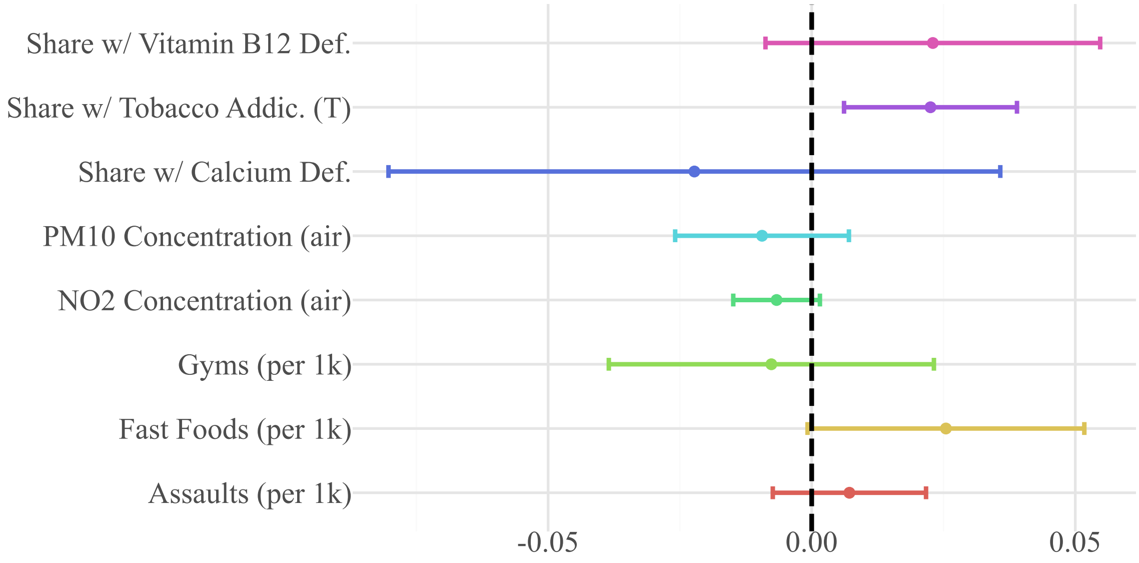

Figure 7: DDML Estimates, Multiple Variables

We find a clear and statistically significant positive effect of the share of the hospital population diagnosed with tobacco addiction on T1D incidence, significant at the 1% level.[4] When analyzing by sex, the estimates become noisier: the effect remains significant for males (at the 10% level) but is not significant for females. While we cannot interpret tobacco use as an individual-level risk factor, due to the risk of ecological fallacy, the robustness of this finding across multiple models and subpopulations supports the conclusion that tobacco consumption at the population level increases the risk of T1D, particularly among males. Additionally, the number of fast food establishments is marginally significant, falling just below the 5% threshold. This result contributes to the growing body of evidence suggesting that unhealthy consumption behaviors within a population are associated with increased T1D incidence.

No other variables reach our chosen level of statistical significance (10%). Notably, the previously observed counterintuitive negative association with NO_2 concentrations is no longer statistically significant. Other environmental variables, such as insecurity, proxied by the number of assaults, are also not statistically significant. However, we did not exclude other crime-related variables from the model, this is a limitation of our use of DML. For some variables we don't have clear enough causal relationship to test and as such we may suffer from collider bias.

[4] (T): Total Population; (M): Males; (F): Females

6 Conclusion

This study aimed to investigate the regional variations in T1D incidence across metropolitan France by examining associations with a comprehensive set of environmental, demographic, and socioeconomic indicators derived from PMSI hospital data and diverse external sources. Adopting an inferential approach, our primary objective was to identify potential risk and protective factors contributing to these geographical disparities.

Our analyses, employing Poisson/Negative Binomial regressions, Fixed Effects models, and Double Machine Learning techniques, yielded several notable findings. We found consistent evidence supporting a protective association between increased sunshine duration (a proxy for Vitamin D synthesis) and lower monthly T1D incidence rates at the department level. This association held significance even when accounting for regional differences in dietary Vitamin D intake. Conversely, a higher prevalence of diagnosed tobacco addiction within a region was robustly associated with increased T1D incidence, an effect particularly pronounced among males. Socioeconomic factors also emerged as relevant, with indicators reflecting lower-to-middle educational attainment showing a positive association with incidence, again more strongly in males, while higher education and income were negatively associated.

It is crucial to interpret these findings within the context of the study's limitations. As an exercise in ecological inference based on aggregate data, our results reflect associations at the regional level and cannot establish individual-level causality. The reliance on a single year of data (2023) restricts longitudinal insights, and the use of proxy variables, such as ketoacidosis diagnoses (E101) for incidence and hospital data for population-level health indicators like Vitamin D deficiency,carries inherent uncertainties. Unobserved confounding factors common to observational studies may also influence the estimates.

Despite these caveats, this research provides valuable insights into potential drivers of T1D heterogeneity in France and highlights promising avenues for future investigation. Leveraging additionnal years of PMSI data would enable more robust longitudinal analyses and refine incidence estimation. Further research should also aim to incorporate individual mobility patterns and utilize more direct measures for environmental exposures, dietary habits, and socioeconomic status.

7 Bibliography

Buchmann, Maike et al. (2023). "Inzidenz, Prvalenz und Versorgung von Typ-1-Diabetes bei Kindern und Jugendlichen in Deutschland: Zeittrends und sozialrumliche Lage (engl.)" In: in collab. with Robert Koch-Institut. Publisher: Robert Koch-Institut. DOI: 10.25646/11439.

Thuillier, Philippe and Jacques Mansourati (Feb. 2023). "Quels sont les liens entre tabagisme et insulinorsistance, insulinosensibilit ?" In: *Mdecine des Maladies Mtaboliques*, S1957255723000202. ISSN: 19572557. DOI: 10.1016/j.mmm.2023.01.006.

Taha-Khalde, Alaa et al. (Sept. 2021). "Air pollution and meteorological conditions during gestation and type 1 diabetes in offspring". In: *Environment International* 154, p. 106546. ISSN: 01604120. DOI: 10.1016/j.envint.2021.106546.

Elten, Michael et al. (May 2020). "Ambient air pollution and incidence of early-onset paediatric type 1 diabetes: A retrospective population-based cohort study". In: *Environmental Research* 184, p. 109291. ISSN: 00139351. DOI: 10.1016/j.envres.2020.109291.

Michalska, Magorzata et al. (Jan. 2020). "Gaseous Pollutants and Particulate Matter (PM) in Ambient Air and the Number of New Cases of Type 1 Diabetes in Children and Adolescents in the Pomeranian Voivodeship, Poland". In: *BioMed Research International* 2020.1. Ed. by Fei He. Number: 1, p. 1648264. ISSN: 2314-6133, 2314-6141. DOI: 10.1155/2020/1648264.

Traversi, Deborah et al. (Oct. 16, 2020). "Risk factors for type 1 diabetes, including environmental, behavioural and gut microbial factors: a casecontrol study". In: *Scientific Reports* 10.1. Number: 1, p. 17566. ISSN: 2045-2322. DOI: 10.1038/s41598-020-74678-6.

Chernozhukov, Victor et al. (Feb. 1, 2018). "Double/debiased machine learning for treatment and structural parameters". In: *The Econometrics Journal* 21.1. Number: 1, pp. C1–C68. ISSN: 1368-4221, 1368-423X. DOI: 10.1111/ectj.12097.

Virtanen, Suvi M (July 2016). "Dietary factors in the development of type 1 diabetes: Dietary factors and risk of type 1 diabetes". In: *Pediatric Diabetes* 17, pp. 49–55. ISSN: 1399543X. DOI: 10.1111/pedi.12341.

Morimoto, Akiko et al. (May 2013). "Impact of cigarette smoking on impaired insulin secretion and insulin resistance in <span style="font-variant:small-caps;">J</span> apanese men: The <span style="font-variant:small-caps;">S</span> aku <span style="font-variant:small-caps;">S</span> tudy". In: *Journal of Diabetes Investigation* 4.3. Number: 3, pp. 274–280. ISSN: 2040-1116, 2040-1124. DOI: 10.1111/jdi.12019.

Vialettes, B. and B. Conte-Devolx (Sept. 2013). “Le diabete de type 1 post-traumatique existe-t-il ?” In: *Mdecine des Maladies Mtaboliques* 7.4. Number: 4, pp. 379–384. ISSN: 19572557. DOI: 10.1016/S1957-2557(13)70606-0.

Zung, Amnon et al. (June 2012). “Increase in the incidence of type 1 diabetes in Israeli children following the Second Lebanon War: Type 1 diabetes following Second Lebanon War”. In: *Pediatric Diabetes* 13.4. Number: 4, pp. 326–333. ISSN: 1399543X. DOI: 10.1111/j.1399-5448.2011.00838.x.

Virk, Jasveer et al. (July 9, 2010). “Early Life Disease Programming during the Preconception and Prenatal Period: Making the Link between Stressful Life Events and Type-1 Diabetes”. In: *PLoS ONE* 5.7. Ed. by Luis Huicho. Number: 7, e11523. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0011523.

Mohr, S. B. et al. (Aug. 2008). “The association between ultraviolet B irradiance, vitamin D status and incidence rates of type 1 diabetes in 51 regions worldwide”. In: *Diabetologia* 51.8. Number: 8, pp. 1391–1398. ISSN: 0012-186X, 1432-0428. DOI: 10.1007/s00125-008-1061-5.

Norman, Anthony W (Aug. 2008). “From vitamin D to hormone D: fundamentals of the vitamin D endocrine system essential for good health”. In: *The American Journal of Clinical Nutrition* 88.2. Number: 2, 491S–499S. ISSN: 00029165. DOI: 10.1093/ajcn/88.2.491S.

Zipitis, C S and A K Akobeng (June 1, 2008). “Vitamin D supplementation in early childhood and risk of type 1 diabetes: a systematic review and meta-analysis”. In: *Archives of Disease in Childhood* 93.6. Number: 6, pp. 512–517. ISSN: 0003-9888, 1468-2044. DOI: 10.1136/adc.2007.128579.

Du Prel, J.-B. et al. (Apr. 2007). “Socioeconomic conditions and type 1 diabetes in childhood in North RhineWestphalia, Germany”. In: *Diabetologia* 50.4. Number: 4, pp. 720–728. ISSN: 0012-186X, 1432-0428. DOI: 10.1007/s00125-007-0592-5.

Holick, Michael F. (July 19, 2007). “Vitamin D Deficiency”. In: *New England Journal of Medicine* 357.3. Number: 3, pp. 266–281. ISSN: 0028-4793, 1533-4406. DOI: 10.1056/NEJMra070553.

Hathout, Eba H et al. (Apr. 2006). “Air pollution and type 1 diabetes in children”. In: *Pediatric Diabetes* 7.2. Number: 2, pp. 81–87. ISSN: 1399-543X, 1399-5448. DOI: 10.1111/j.1399-543X.2006.00150.x.

Pundzit-Lycka, Aust et al. (Dec. 1, 2004). “Diet, Growth, and the Risk for Type 1 Diabetes in Childhood”. In: *Diabetes Care* 27.12. Number: 12, pp. 2784–2789. ISSN: 0149-5992, 1935-5548. DOI: 10.2337/diacare.27.12.2784.

Hyppnen, Elina et al. (Nov. 2001). “Intake of vitamin D and risk of type 1 diabetes: a birth-cohort study”. In: *The Lancet* 358.9292. Number: 9292, pp. 1500–1503. ISSN: 01406736. DOI: 10.1016/S0140-6736(01)06580-1.

Littorin, Bengt et al. (June 1, 2001). “Family Characteristics and Life Events Before the Onset of Autoimmune Type 1 Diabetes in Young Adults”. In: *Diabetes Care* 24.6. Number: 6, pp. 1033–1037. ISSN: 0149-5992, 1935-5548. DOI: 10.2337/diacare.24.6.1033.

Ostgren, C. J. et al. (June 2000). “Associations between smoking and cell function in a non-hypertensive and nondiabetic populationSkaraborg Hypertension and Diabetes Project”. In: *Diabetic Medicine* 17.6. Number: 6, pp. 445–450. ISSN: 0742-3071, 1464-5491. DOI: 10.1046/j.1464-5491.2000.00294.x.

Not Available, Not Available (Jan. 11, 1999). “Vitamin D supplement in early childhood and risk for Type I (insulin-dependent) diabetes mellitus”. In: *Diabetologia* 42.1. Number: 1, pp. 51–54. ISSN: 0012-186X, 1432-0428. DOI: 10.1007/s001250051112.

A Appendix

A.1 Principal Component Analysis

A.1.1 Main Loadings From Full PCA (ARR)

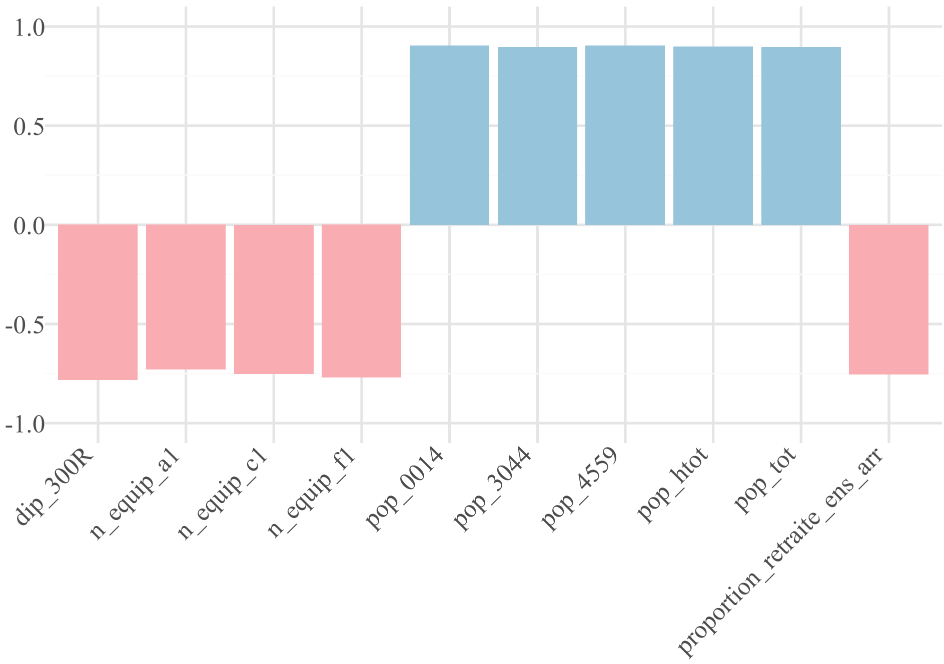

Figure 8: Top 5 Positive and negative variables PC1

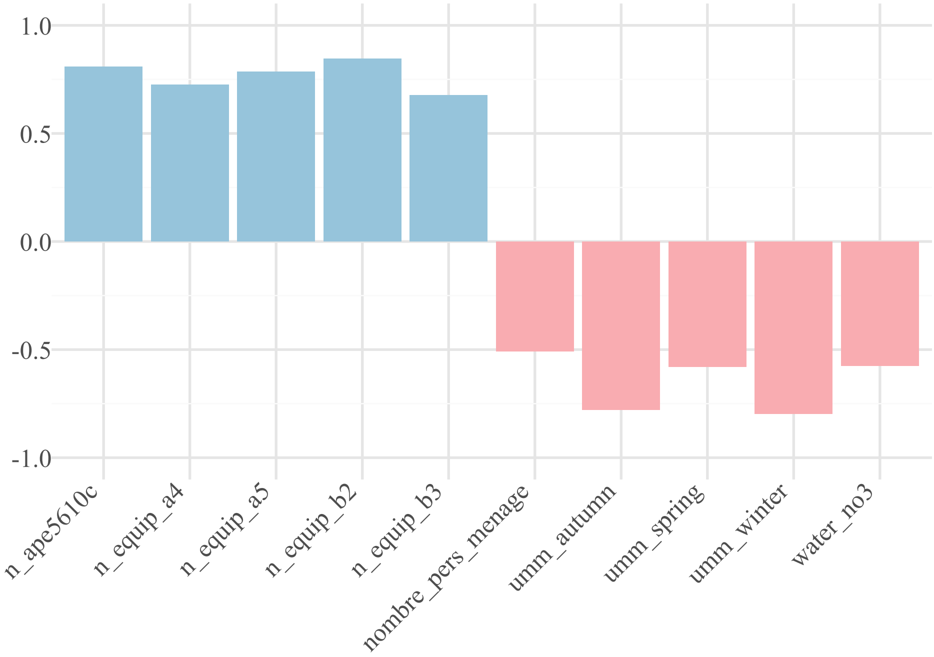

Figure 9: Top 5 Positive and negative variables PC2

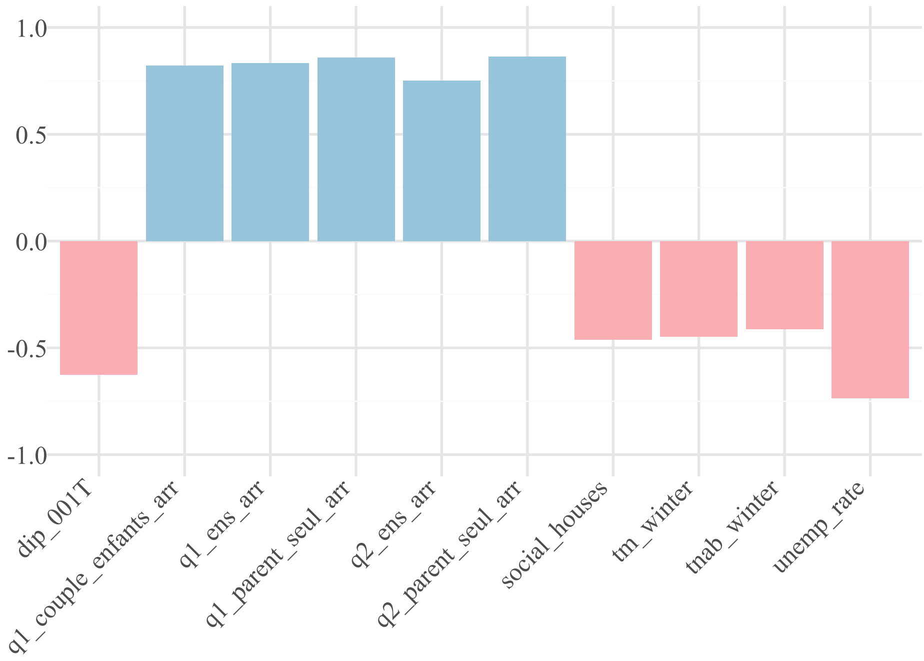

Figure 10: Top 5 Positive and negative variables PC3

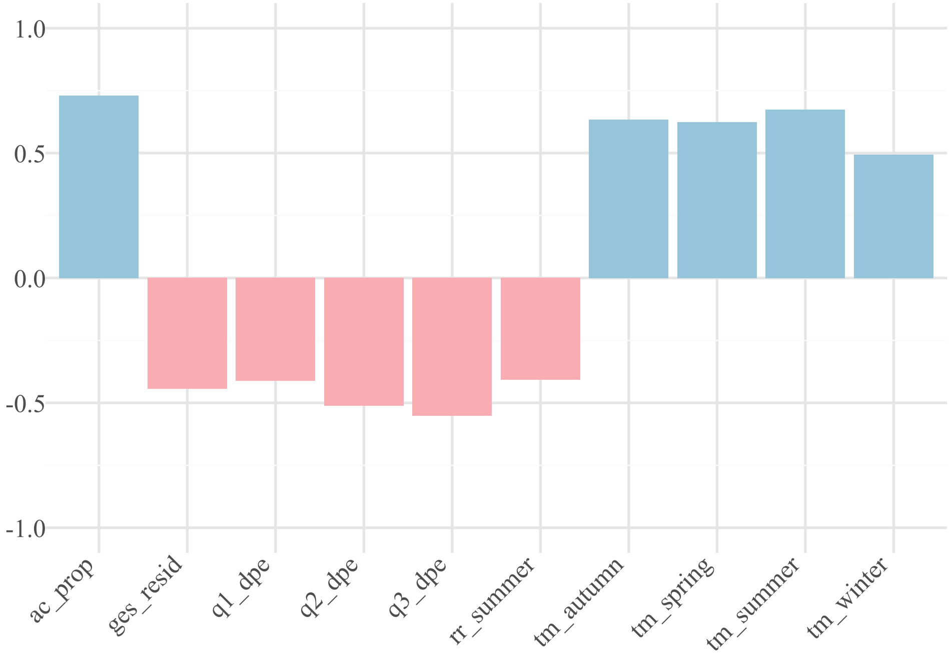

Figure 11: Top 5 Positive and negative variables PC4

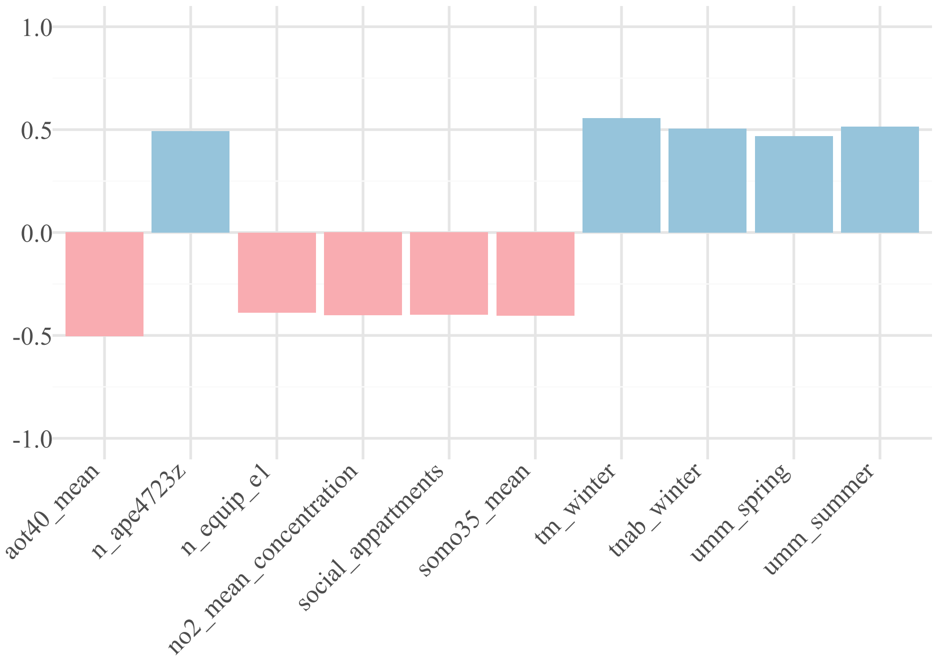

Figure 12: Top 5 Positive and negative variables PC5

A.1.2 Cartography of France Full PCA

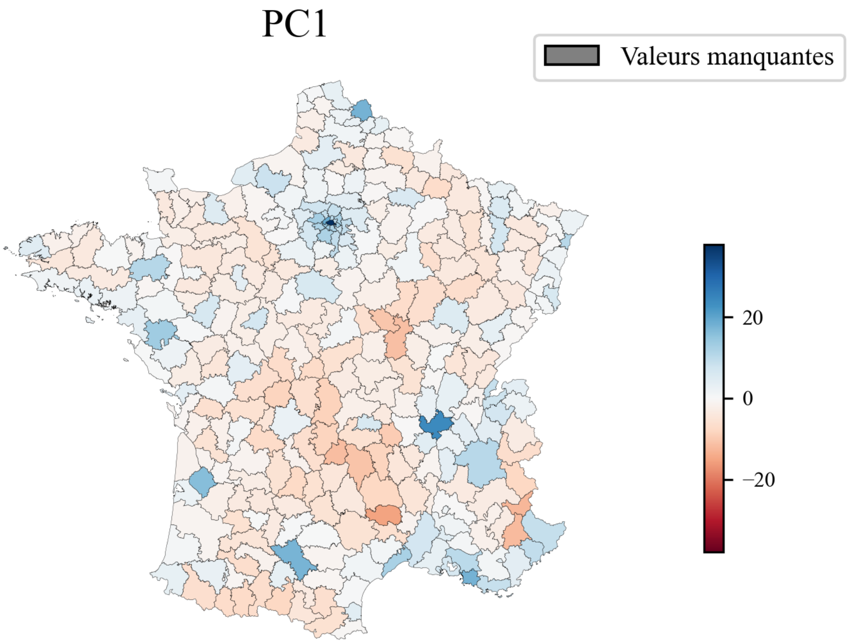

Figure 13: Geographic Representation PC1

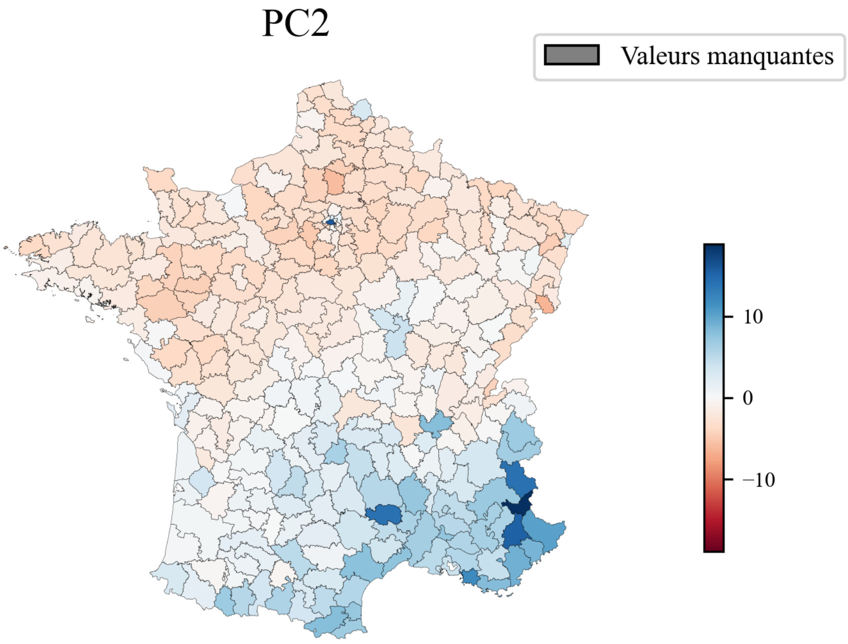

Figure 14: Geographic Representation PC2

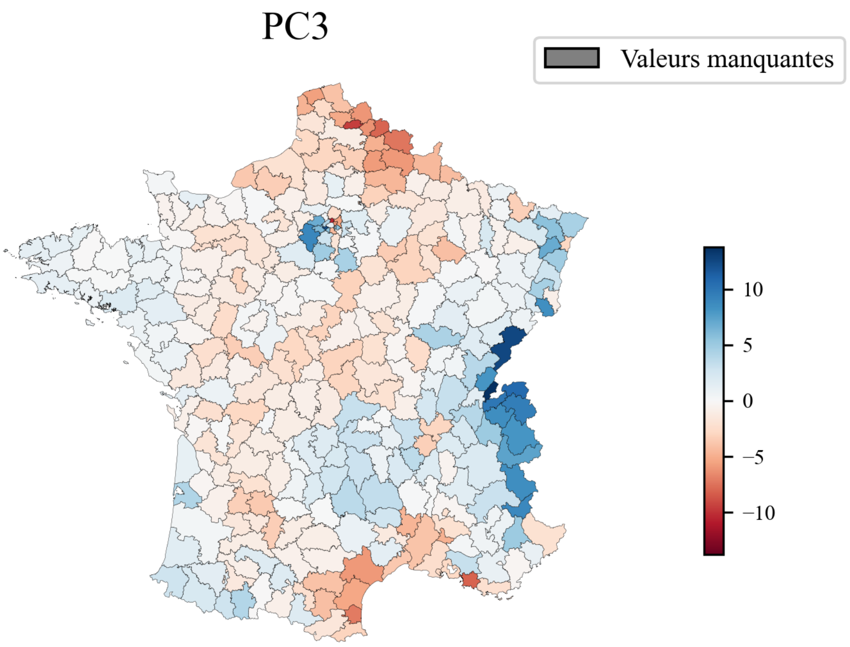

Figure 15: Geographic Representation PC3

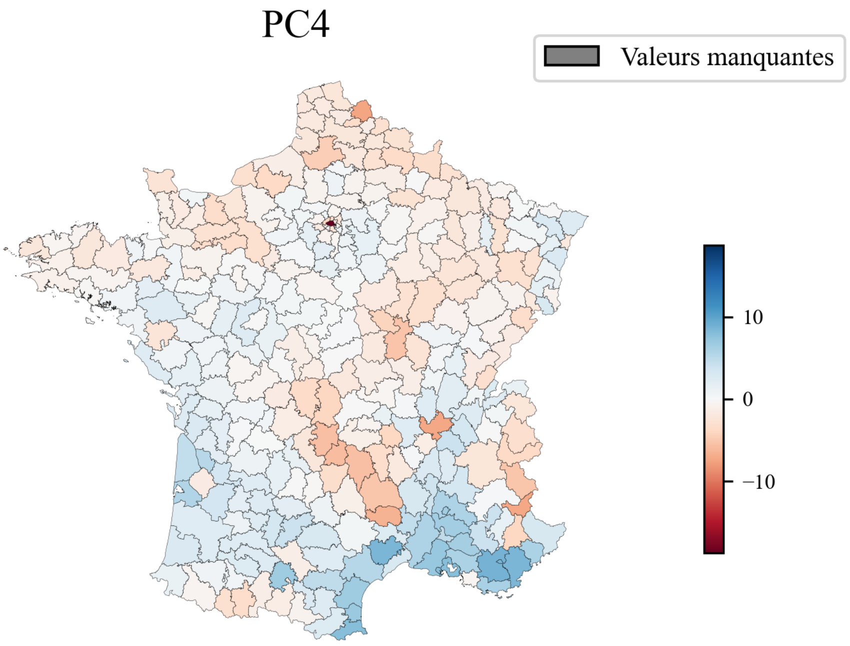

Figure 16: Geographic Representation PC4

A.1.3 Variable loadings from PCA-SES (ARR)

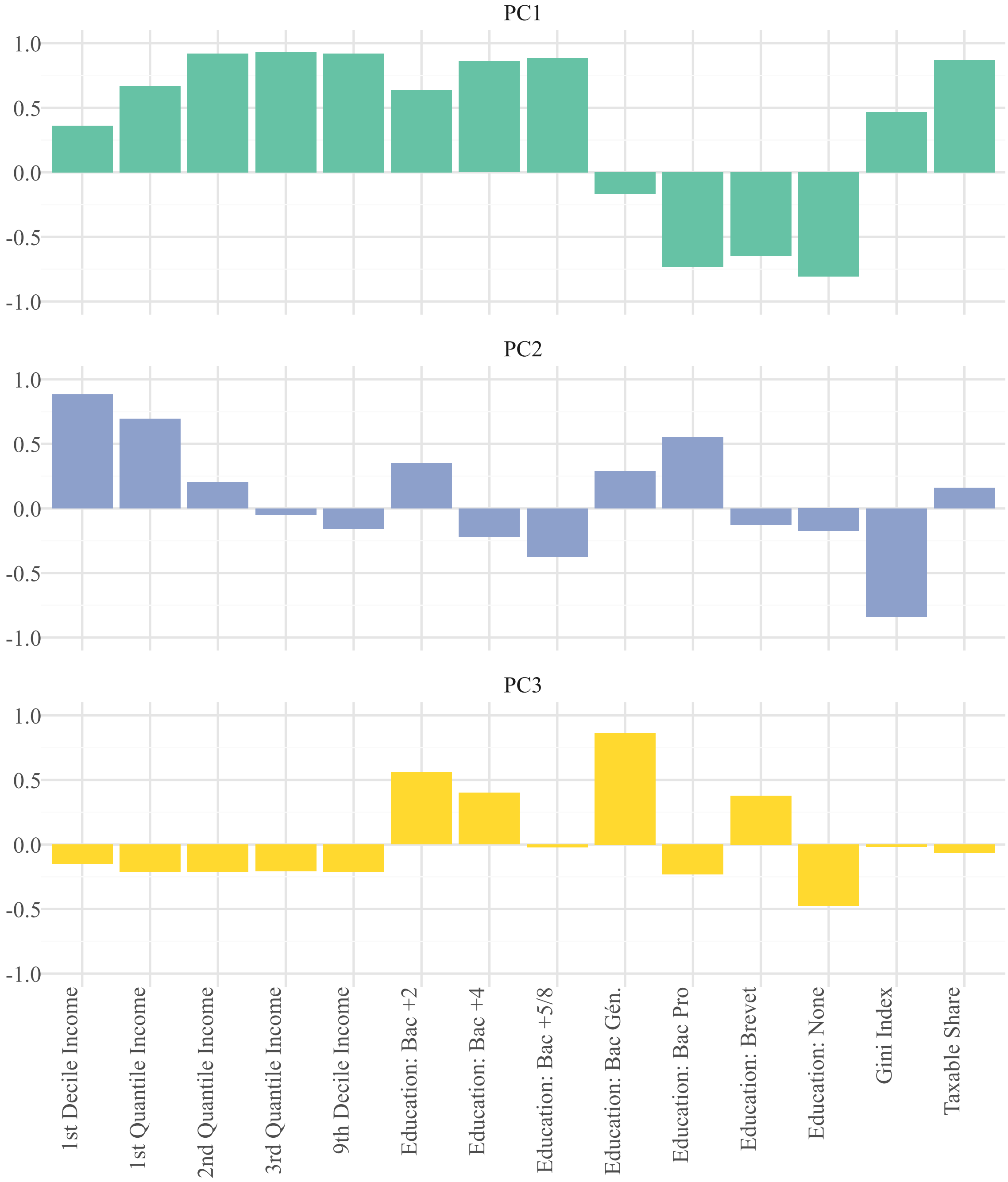

Figure 17: Feature Loading PCA-SES (ARR)

A.1.4 Variable loadings from PCA-SES (BV2022)

We plot the incidence and prevalence rates over French departments to see a priori relations with our main axis of variation. Such visualisation are useful for variable selection see the next appendix.

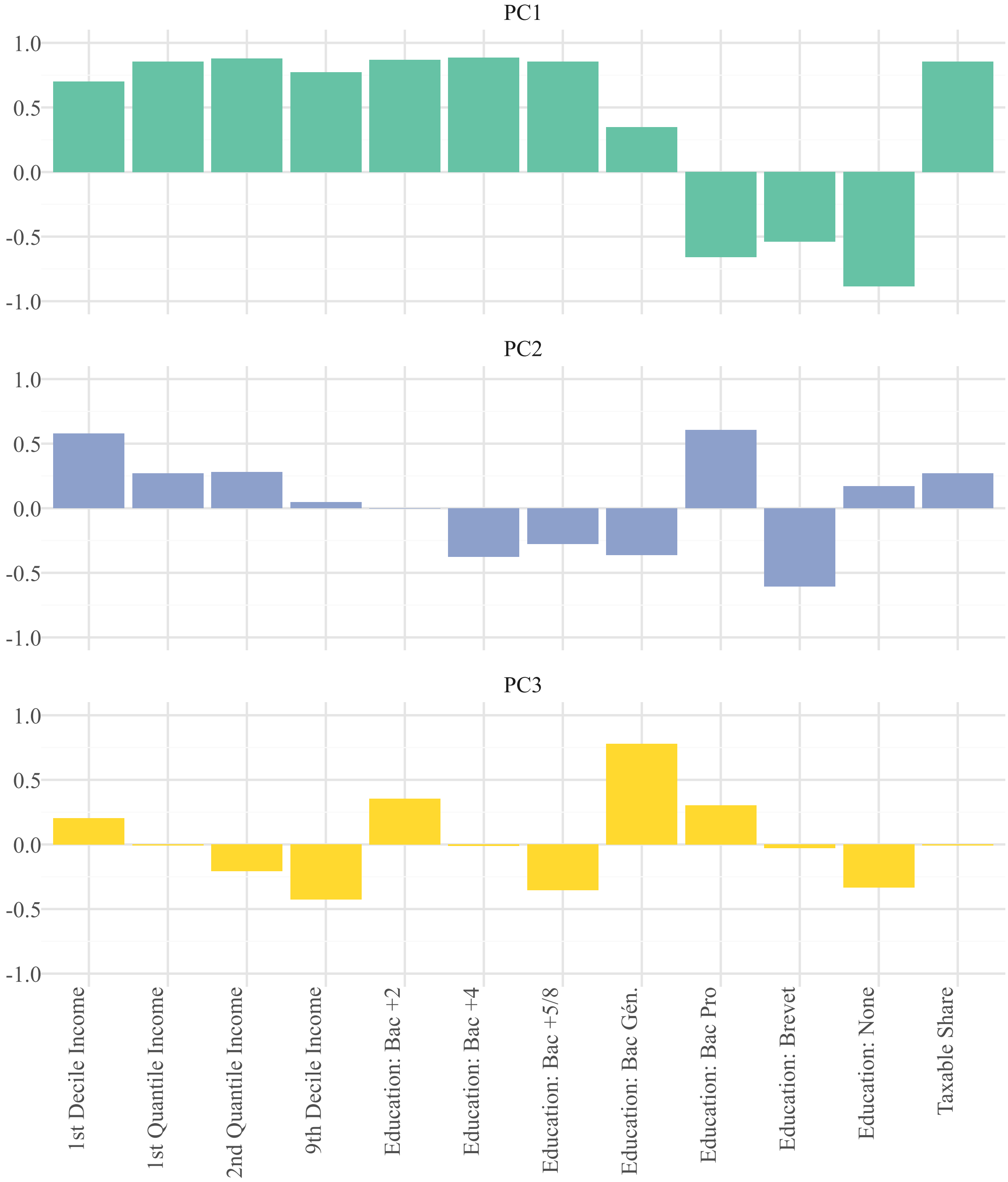

Figure 18: Feature Loading PCA-SES (BV)

A.2 Health Cartography

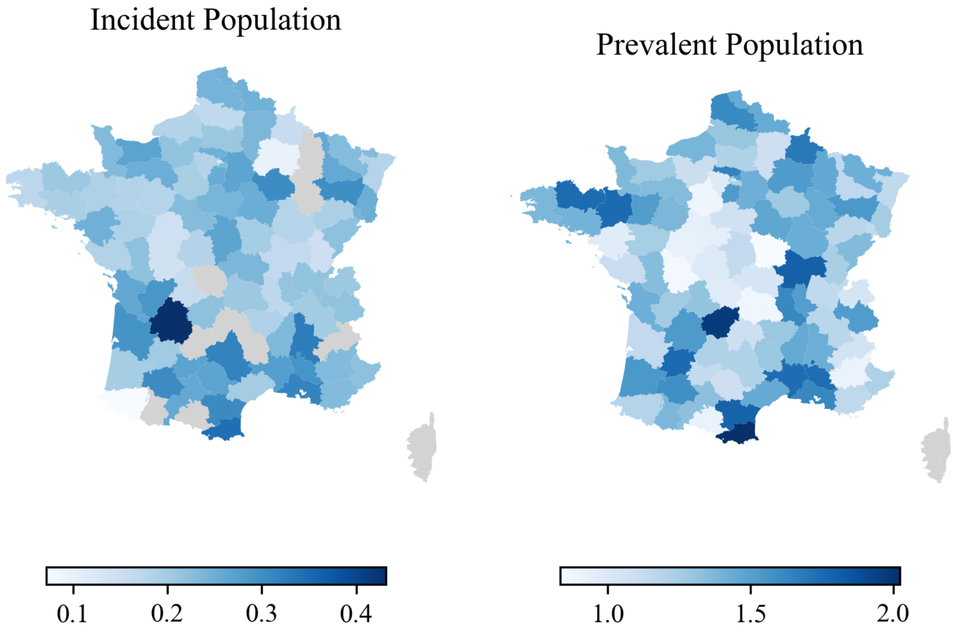

Figure 19: T1D in France, 2023, per 1,000

A.3 Selected variables

Name	Type	Meaning	Justification / Hypothesis
Socio-economic PC1	C	First principal component of the PCA on SES variables	Used to C for SES
Socio-economic PC2	C	Second principal component of the PCA on SES variables	Used to C for SES
Socio-economic PC3	C	Third principal component of the PCA on SES variables	Used to C for SES
Summer temperature	T	Average temperature during summer 2023	Ts for physiological stress from heat exposure
Winter temperature	T	Average temperature during winter 2023	Ts for influence of moderate cold exposure
Excess Ozone	T	Annual mean of SOMO35 (ozone exposure)	Ts air quality impact
PM10 Concentration	T	Mean PM10 concentration (µg/m^3)	Ts air quality impact
O3 Concentration	T	Mean O_3 concentration (µg/m^3)	Ts air quality impact
NO2 Concentration	T	Mean NO_2 concentration (µg/m^3)	Ts air quality impact
Residential GHG	T	Residential greenhouse gas emissions	Proxy for environmental quality
NO3 Concentrations	T	NO_3 concentration in drinking water	Cled by *prop_robinet*; Ts water quality
Water pH	T	pH level of drinking water	Cled by *prop_robinet*; Ts water quality
Medical Facilities	C	Medical/paramedical professionals (private)	Potential influence on early T1D detection
Public Services	C	Local public services	Socio-environmental infrastructure C

Continued on next page

Name	Type	Meaning	Justification / Hypothesis
Teaching Primary	C	Primary education institutions	SES / public resource C
Teaching Secondary	C	Lower secondary education institutions	SES / public resource C
Fast Foods	T	Fast food outlets	Ts hypothesis of diet-related risk for T1D
Gyms	T	Sports clubs	Ts access to physical activity
Butchers	T	Meat retailers	Ts dietary access hypothesis
Fishmongers	T	Fishmongers	Linked to vitamin D intake
Assaults	T	Assault complaints per inhabitant	Proxy for ambient stress and safety
Share of Homes with AC_prop	T	Share of households with AC	Proxy for heatwave protection
Share of tap drinkers	C	Proportion drinking tap vs bottled water	Used to interpret water-related quality variables
Share w/ Vitamin D Def.	T	Diagnosed vitamin D deficiency	Vitamin D and T1D link (with reverse causality caution)
Share w/ Calcium Def.	T	Diagnosed calcium deficiency	Nutritional deficiency hypothesis (reverse causality risk)
Share w/ Family Issues	T	Indicator of family-related stress	Psychosocial stress hypothesis
Share w/ Mineral Def.	T	Diagnosed mineral deficiency	General mineral deficiency and T1D risk
Share w/ Tobacco Addic.	T	Indicator of tobacco use	Lifestyle-related trigger for T1D onset
Share w/ Vitamin B9 Def.	T	Diagnosed vitamin B9 deficiency	Nutritional deficiency hypothesis
Share w/ Vitamin B12 Def.	T	Diagnosed vitamin B12 deficiency	Nutritional deficiency hypothesis

A.4 Double ML Presentation

Random forest is an ensemble learning method for classification or regression that leverages creation of a multitude of decision trees during training. Indeed, a single decision tree can easily overfit the training data. Random Forest overcomes this by constructing a forest of trees, each trained on a random subset of the data. Each tree makes a prediction and for regression tasks the Random Forest output is the average prediction of the different trees. This method is non-parametric, it does not assume any functional form between the covariates and the outcome. It can handle non-linear relationships, interactions between variables and high-dimensional feature spaces.

The training algorithm for random forests applies the general technique of bootstrap aggregating or bagging, to tree learners. With a training set (X, Y), bagging T times consist in selecting a random sample with replacement of the training set and fits trees to these samples. The samples are named (X_t, Y_t) for $t = 1, ..., T$. The second step is train the regression tree f_t on X_t, Y_t. After training, the forest aggregates over the diverse trees:

$$\hat{f}(x) = \frac{1}{T}\sum_{t=1}^{T} f_t(x)$$

where $f_t(x)$ is the prediction from the t-th tree.

We used random forests to estimate the nuisance components within the Double Machine Learning procedure, which allows to learn complex patterns in the data without over-fitting. The Double/Debiased Machine Learning (DDML) framework is presented in the next paragraph. \

Double/Debiased Machine Learning, also called de-biased machine learning, was first introduced by Chernozhukov et al. 2018. The goal is to develop a causal estimator by leveraging the flexibility of non-parametric machine learning methods. The goal of this method is also to reduce bias, provide a valid confidence interval and a 'root-n-consistent' estimator (e.g. estimator error approaches zero at a rate of $1/\sqrt{n}$ when the sample size n goes to infinity).

The goal is to isolate the causal effect of a treatment variable D on an outcome Y (T1D incidence) while controlling for a set of confounding variables X. The DML procedure relies on the **Frisch-Waugh-Lovell (FWL) theorem**, which states that in a linear model, the coefficient of a regressor D in the full regression of Y on D and X is equal to the coefficient from regressing the residuals of Y (after removing the effect of X) on the residuals of D (after removing the effect of X).

This motivates the three-step structure of DML:

1. Estimate nuisance functions using machine learning:

$$g_0(X) = E[Y|X] \qquad m_0(X) = E[D|X]$$

2. Partialling out the confounders:

$$\tilde{Y} = Y - g_0(X) \qquad \tilde{D} = D - m_0(X)$$

   This second step removes variation associated with confounders.

3. Estimate the causal effects with residuals:

$$\tilde{Y} = \theta_0 \tilde{D} + \zeta$$

   The θ_0 coefficient is an estimate of the causal impact of the treatment.

The estimator is orthogonal, meaning that estimation error in nuisance functions doesn't bias $\hat{\beta}$, asymptotically normal and double-robust (e.g. is consistent as long as the nuisance components are estimated consistently). The combination of Random Forest and DML offers flexibility without increasing bias: non-parametric machine learning methods can capture complex patterns in data while DML corrects for the lack of causal interpretability. It also offers robustness because by removing the influence of confounders and using residuals, the estimator is less sensitive to model misspecification and overfitting.

A.5 Full Regression Results (Poisson / Neg. Binomial)

Below are the results of the main Poisson and Negative Binomial regressions with all coefficients and associated p-values reported. The model numbering follows the one described in the report.

Table 5: Main Regression Results (Full)

	(1)	(2)	(3)	(4)
Intercept	0.00079***	0.00025***	0.00039***	0.00023***
	0.0074	0.0	0.0005	0.0
Socio-economic PC1	1.0011	0.96434***	1.00276	0.96524***
	0.9262	0.0015	0.7813	0.0003
Socio-economic PC2	0.99695	0.9514**	1.00218	0.96168*
	0.8868	0.0337	0.9051	0.0579
Socio-economic PC3	1.06621***	1.00051	1.07187***	0.99136
	0.0013	0.9804	0.0	0.6273
Share of Homes with AC	2.69198*	1.55551	2.19283	1.2701
	0.0978	0.1855	0.1273	0.4143
Assaults (per 1k)	1.05691**	1.04763	1.06043***	1.0487
	0.0167	0.2587	0.0004	0.2093
Share w/ Calcium Def.	1.08415	1.37205**	1.09553	1.35973***
	0.4398	0.0155	0.314	0.0078
Share w/ Family Issues	0.99394	0.9991	0.99975	1.01625
	0.8783	0.9759	0.9943	0.5394
Share w/ Mineral Def.	0.88198	0.93158	0.90172	0.98171
	0.2049	0.662	0.2534	0.8969
Share w/ Tobacco Addic. (T)	1.07036***	1.03436*	1.07237***	1.03409**
	0.0077	0.0724	0.002	0.0475
Share w/ Vitamin B12 Def.	1.10438	1.08613	1.10653	1.09703
	0.1376	0.3471	0.1049	0.2426
Share w/ Vitamin B9 Def.	0.92666	1.02621	0.93504	1.03272
	0.3168	0.7384	0.3395	0.6506

Continued on next page

	(1)	(2)	(3)	(4)
Share w/ Vitamin D Def.	1.09732	1.03616	1.11687*	1.03458
	0.1988	0.4053	0.0681	0.3697
Residential GHG (per 1k)	0.99996	1.00006	0.99999	1.00001
	0.7736	0.5638	0.9525	0.9413
Butchers (per 1k)	1.84047**	1.05507	1.73818**	1.09682
	0.0349	0.7435	0.0333	0.5452
Fishmongers (per 1k)	0.93924	0.80079	0.95952	0.72908
	0.8798	0.3495	0.9086	0.1584
Fast Foods (per 1k)	1.02338	1.02698	1.00688	1.04175
	0.6495	0.3977	0.8706	0.1584
Gyms (per 1k)	0.94446	0.94531*	0.94451	0.93976**
	0.1916	0.0542	0.1369	0.0183
Public Services (per 1k)	0.88718**	0.90435***	0.88847***	0.88466***
	0.0149	0.006	0.009	0.0004
Teaching Primary (per 1k)	1.07936	0.98173	1.07469	1.00353
	0.6519	0.8728	0.638	0.974
Teaching Secondary (per 1k)	0.86373	1.27365	1.03835	1.4022
	0.8976	0.6254	0.9711	0.4767
Medical Facilities (per 1k)	0.93206***	0.99154	0.92538***	0.99222
	0.0004	0.595	0.0	0.5467
NO2 Concentration (air)	0.9781*	0.99084	0.97804**	0.98052**
	0.0891	0.4405	0.0427	0.0285
O3 Concentration (air)	1.00416	0.99824	1.00349	0.99595
	0.6952	0.8184	0.7006	0.51
PM10 Concentration (air)	1.0153	1.00138	1.02227	1.01882
	0.5584	0.9439	0.3098	0.2248
Share of tap drinkers	0.97625	-	0.97813	-
	0.3959	-	0.349	-
Tap drinking x Water PH	1.00309	-	1.0029	-

Continued on next page

	(1)	(2)	(3)	(4)
	0.401	-	0.3435	-
Excess Ozone (air)	0.99993	1.00002	0.99994	1.00003
	0.1443	0.5623	0.1668	0.258
Summer temperature	1.02686	-	1.03376	-
	0.4183	-	0.2383	-
Winter temperature	0.95711	-	0.96391	-
	0.1083	-	0.1116	-
NO3 Concentrations (water)	0.99027	-	0.99103*	-
	0.1302	-	0.0903	-
Tap drinking x NO3	1.00004	-	0.99999	-
	0.6622	-	0.9416	-
Water PH	0.83966	-	0.8839	-
	0.5751	-	0.6399	-
AIC	1745.325	4649.565	1753.395	4670.682
DF	282.0	1602.0	282.0	1602.0
Dispersion	-	-	1.513	1.197
Pseudo R^2	0.301	0.314	0.096	0.023

A.6 Common Support Hypothesis (DML)

For our Double ML estimates to be valid, two key conditions must be satisfied:

1. **Common Support:** For every observed value of X_i there must be a positive probability of receiving each level of the treatment variable D.
2. **Correct Model Specification:** The machine learning algorithm must provide an adequate approximation of the true underlying functions

Note that, for proper causal inference, we would also need that: $Y(d) \perp D|X$, that, conditional on X, treatment assignment D is as good as random.

A simple and intuitive way to assess both conditions is to plot the predicted values against the observed values. If the points cluster tightly around the 45 line, this indicates both good predictive performance (suggesting appropriate model specification) and sufficient overlap in the covariate distributions across treatment groups.

In our case many variables fit well this criterion, Assaults, Gyms, NO_2 and Fast Foods are mostly clustered around the 45 line. PM10 concentration and Share of patients with Tobacco addiction exhibit a worse fit, but in the mid region we still have a reasonable predictive accuracy.

In our case, several variables exhibit good fit based on this criterion. In particular, assaults, gyms, NO_2 concentration, and fast food count show predicted values that closely follow the 45 line, suggesting strong model performance and common support. In contrast, PM_{10} concentration and the share of patients diagnosed with tobacco addiction has a weaker fit, particularly at the extremes. Nevertheless, in the mid-range of the predicted values, predictive accuracy remains acceptable.

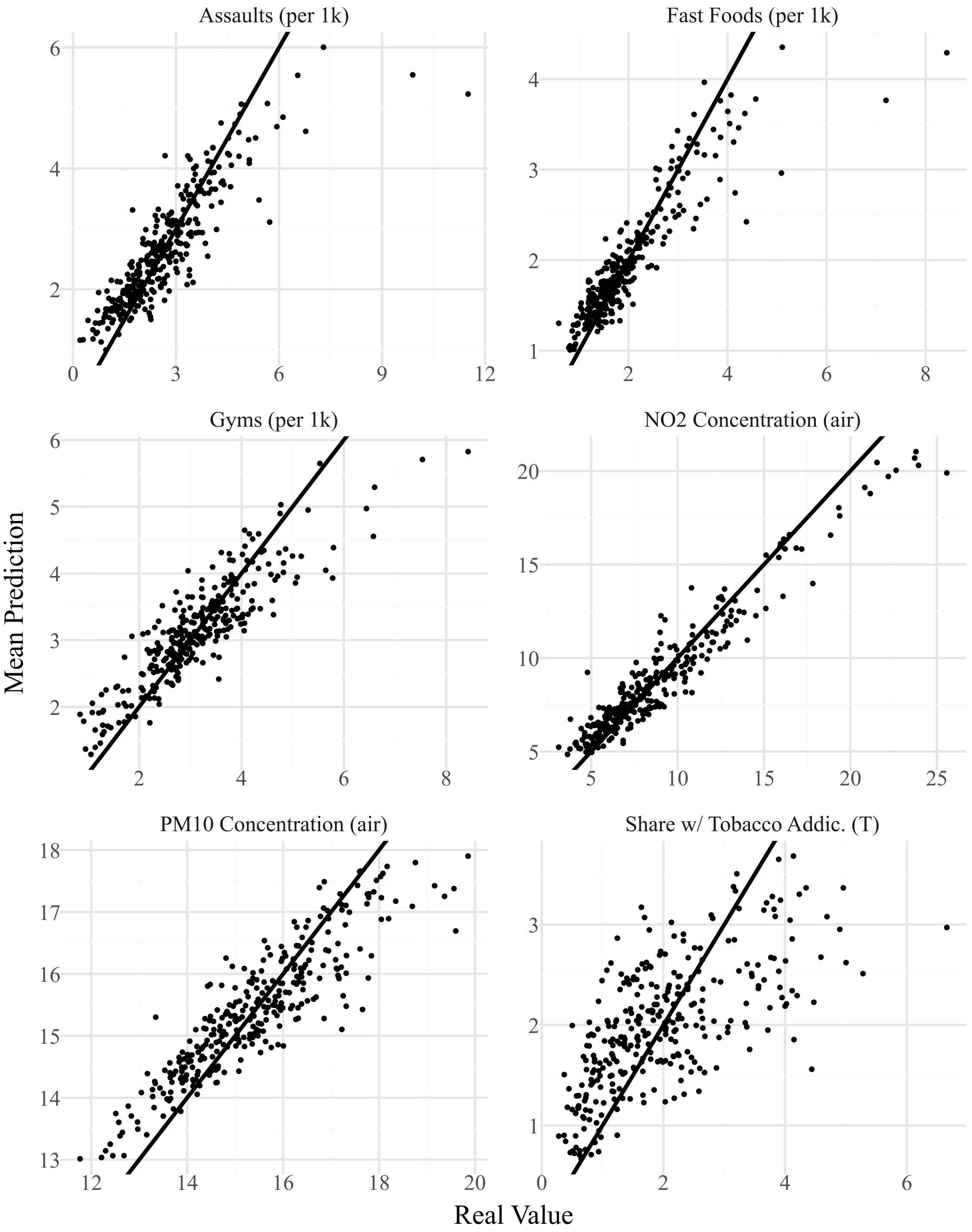

Figure 20: Double ML Residuals

A.7 Poisson FE Table

The table below presents the results from our fixed effects Poisson models.

1. Model (1) includes the entire population.
2. Model (2) is restricted to departments with lower consumption of vitamin D rich foods.
3. Model (3) is fitted on departments with higher consumption of vitamin D rich foods.

P-values are reported below each coefficient. Standard errors (and corresponding p-values) are clustered at the department level to account for intra-departmental correlation.

Table 6: Poisson Fixed Effects

	(1)	(2)	(3)
Humidity (mean)	1.00108	1.00223	1.00134
	0.8011	0.7581	0.8048
Precipitations (mm)	0.99958	0.9988*	1.00012
	0.3338	0.0865	0.8281
Sunshine (days)	0.9621***	0.96125*	0.96763**
	0.0014	0.051	0.0339
Temperature (avg)	1.03192*	1.00733	1.05398**
	0.0535	0.777	0.0186
Temperature (max)	0.99466	1.00562	0.98624
	0.4998	0.6794	0.1786
Temperature (min)	0.97978*	0.99117	0.96855**
	0.0649	0.6008	0.0328
N. Observations	1128.0	540.0	588.0
Dispersion	1.052	1.09	1.017
Pseudo R^2	0.44	0.333	0.484