World Inequalities

Dataset

The data used for this visualization contains inequities and macroeconomics data from different dataset :
World Inequality Database (WID) and the ***World bank***.

The data we use for the visualization are :

- **The gini index** (Gini) : measuring the inequality of the income in a country
- **Purchasing Power** (PP)
- **Gross Domestic Product** per midyear population (GDP) : monetary measure of the market value of all the final goods, measuring the wealth of a country
- **Economic Freedom index** : measuring the degree of economic freedom in the world's nations (Investment Freedom, Trade Freedom)
- **Government Integrity** : Corruption Perceptions Index
- **Property rights** : scores the underlining institutions of a strong property rights regime
- **Unemployment**
- **Inflation By Consumer Price**

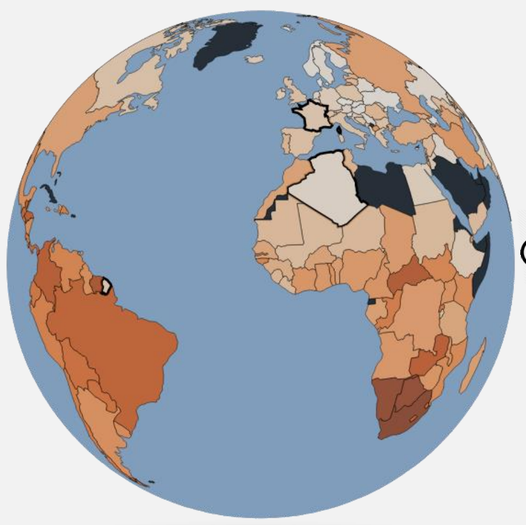
Objectives

The users aimed by this project are **economists wanting** to better **understand inequalities information.**

They are trying to **understand the evolution of inequalities, differences between countries** as well as **the impact of economic and historical events.**

The visualization will be providing tools to communicate the data and have a summary of the inequalities for the economists to exploit in their research

Interactions



World Map

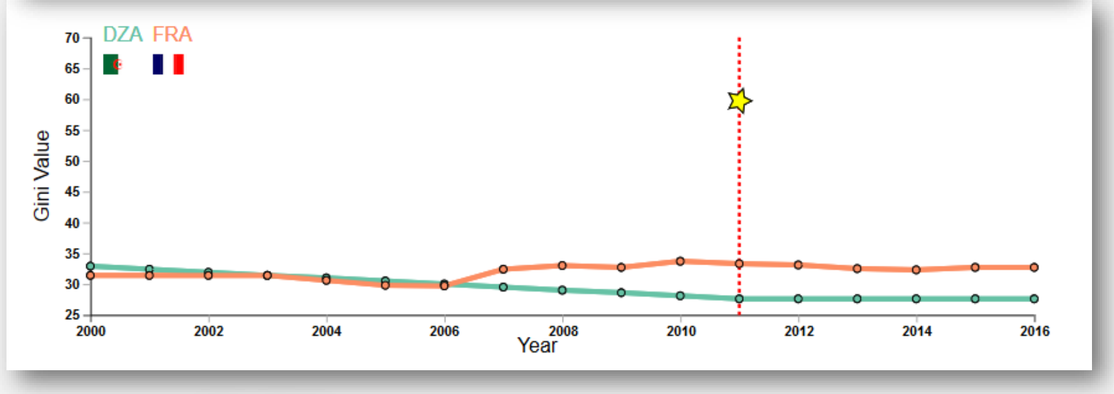
Select country from map

The map is updated by indicator selection and year choice

Time Chart

Choose an indicator

Gini | PP | GDP | Investment freedom | Trade freedom | Government integrity | Property rights



The time chart is updated by indicator selection and country selection

Choose an indicator

Gini | PP | GDP

Reset countries selection

France | Algeria | Russia

Reset

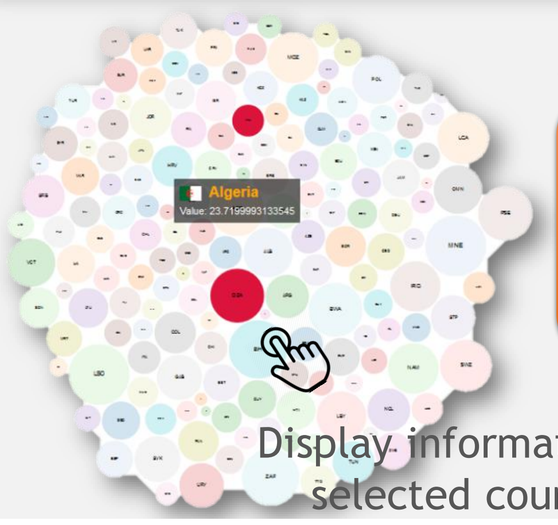
Type a country

2D | 3D

Select country from map

2005 | 2006 | 2007 | 2008 | 2009

Select a year



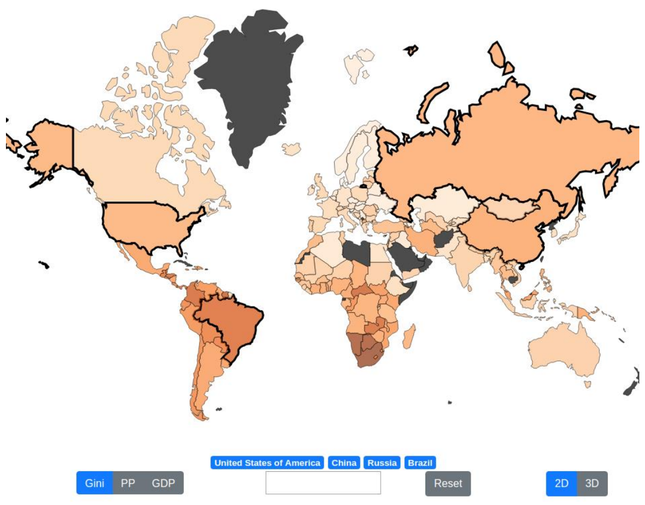
Bubble chart

The bubble of unemployement and inflation is updated by country selection and year choice

Display information of selected country

Observations

1. Exploring the Data



First, we start by moving the year slider and choose the map indicator. You can notice how the indicators evolve through time. For instance, we observe a strong evolution of both GDP and Gini index for Russia and Brazil as well as China and the USA having the highest GDP. Let's select these countries for further analysis.

2. Analyzing chosen Data

Time Chart

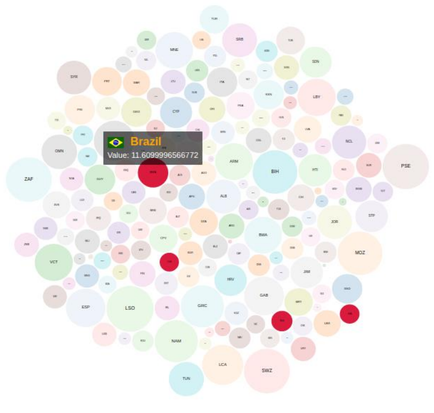
Inequality Indicators Evolution

Here, we can see more attributes and analyze the evolution over the period. By moving the slider, you can also see related events that might explain some of the results. For example, here we see that 2008's financial crisis had a lesser impact on China than it had on other countries. We can also notice that despite having similar GDP curves, Brazil has a lot more inequalities that Russia (cf Gini). Let's concentrate on Russia and Brazil.

3. Going further

Bubble Chart

Unemployement Per Country



Now we have a specific observation we want to check : why has Brazil a much higher Gini index than Russia but the same GDP curve? For that, we can investigate even further by changing representation and going to the unemployment rate visualization. With this, we can see than Brazil has had both a much higher unemployment rate and a higher inflation by consumer price than Russia for the better part of the period, leading to Brazil's higher inequalities.

Road accidents in France

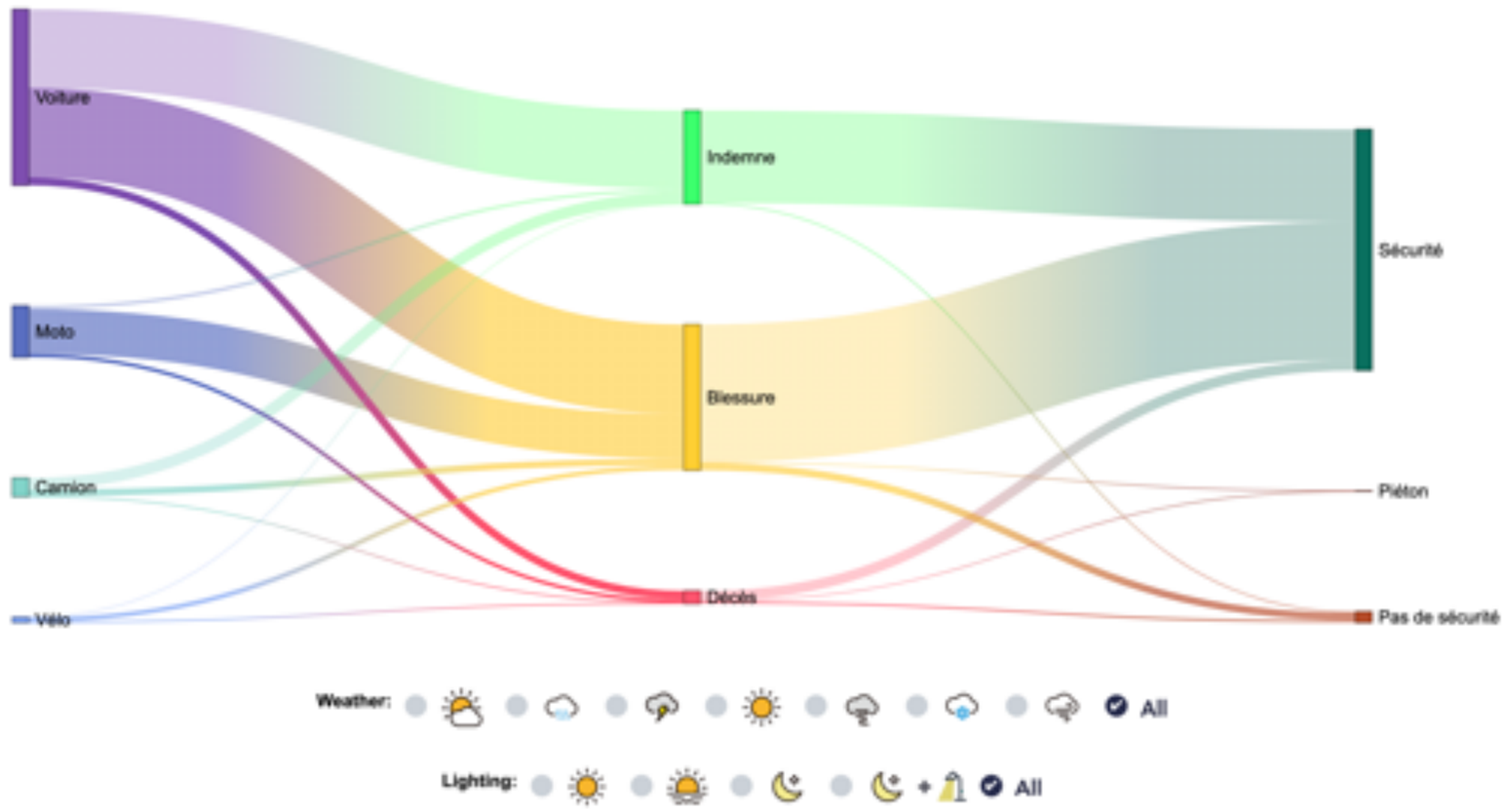
Every year, around **560 billion** kilometers are being travelled by car in France. This is 1.5 million times more than the distance between earth and the moon.

As a direct consequence, yearly, more than 3'500 people die yearly on the French roads. Since 2005, the French government has been collecting data about accidents, and open-sourced the data on data.gouv.fr. There are several public actors in France in charge of the road safety : communes, departments, the state and private companies. These actors are responsible for 3 major objectives :

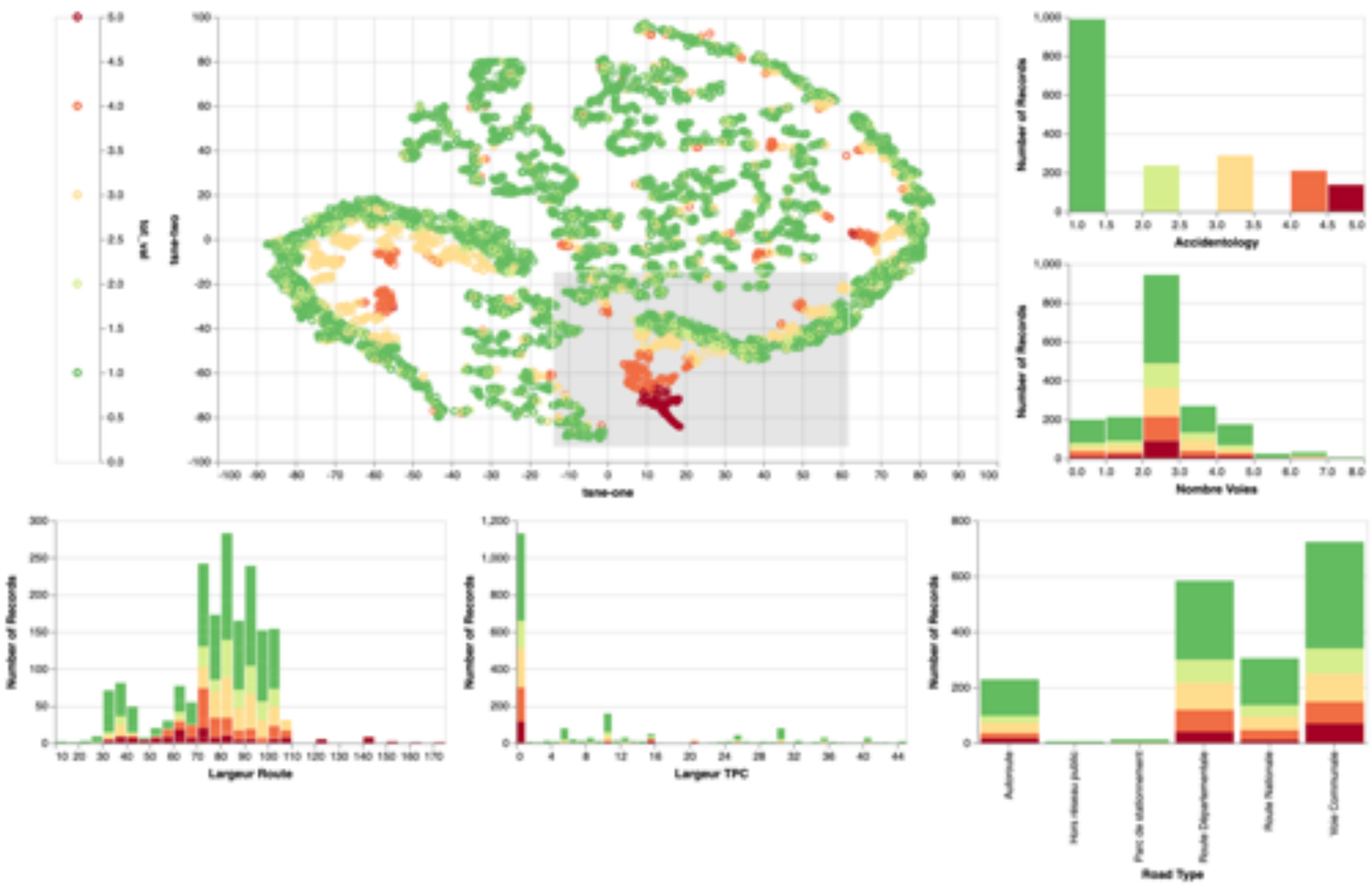
Sensitization

Local police and communes are working together on sensitization of the youngster on the road dangers. If they are involved in an accident, death rates among bike riders and pedestrians are incredibly high compared to other transportation means.

We created a Sankey diagram that shows the proportion of cars, bike riders, motorcycles and trucks involved in a crash, displays the survival rate among each category and distinguishes if they were wearing all security equipments. This graph clearly illustrates the fragility and the exposure of motorcyclists and bike riders, and contributes to sensitization of these populations.



Prevention



Can communes, departments, or even private companies prevent crashes before they occur? To answer this vast question, we created a graph whose role is to cluster the types of roads (width, surface, infrastructure, proximity of a school…), and their related accident rate, thanks to a T-SNE algorithm. We reduce dimension to allow the user to visualize high dimensional problems in a simple dashboard, and understand local distributions of features.

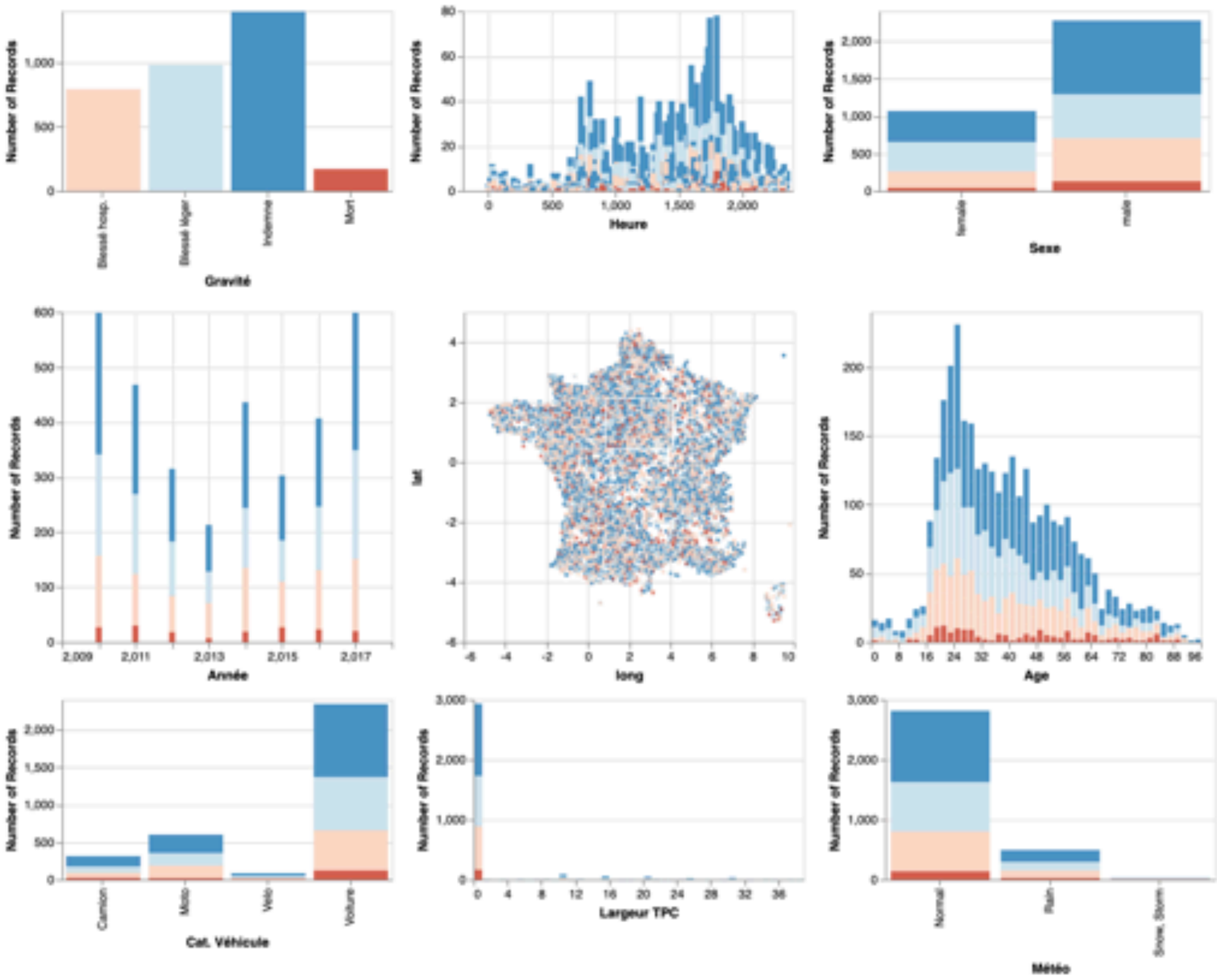
The user can select the type of road from the different pages (communal, departmental, national, highway), and in each case, observe the clusters created by T-SNE. The green clusters represent a low accident rate, and the red ones a high one. The user can hover on a given point to observe all the characteristics of a given road, and select a region in the T-SNE plot. When selecting a region, the histograms all around the plot are updated.

Monitoring

The last task of road services is to monitor the roads and the accident rate on each road. Additional security measures might be needed in certain cases, or certain conditions. Local authorities might have an idea of what type of road is dangerous, but our visualization brings a clear overall view and allows the user to understand a high dimensional spatial problem and the distributions of several features that impact the severity of an accident.

The user selects regions on the map of France, and all the histograms are updated live. This can be useful to understand the distributions in a given region, and by navigating though the different tabs of the website, focus on the different types of roads.

The map of France becomes a spatial trackpad in some sense, and the color scheme corresponds to the severity of the accidents.



Une interface pour explorer les performances des fermes solaires

Introduction

- Nous avons choisi de travailler avec les données de la société **Engie**, à partir du travail de notre projet fil rouge. Ces données représentent les **séries temporelles de la production d'électricité solaire** de la ferme de Blond en Haute-Vienne (Figure 1).
- La ferme solaire est composée d'environ **25 000 panneaux** photovoltaïques, dont la production est regroupée en **8 zones** indépendantes. Les données principales correspondent à :
  - **La production électrique** de chaque zone toutes les minutes de 4h à 20h, du 31 mai 2017 au 4 novembre 2018, comme illustrée lors d'une journée typique sur la Figure 2.
  - **L'irradiance**, qui mesure l'énergie du rayonnement solaire.

Fig. 1: La ferme solaire de Blond

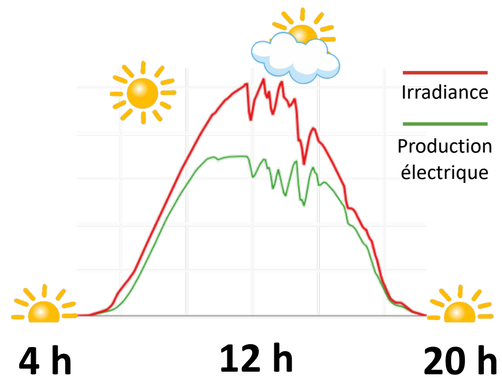


Fig. 2: Une journée typique d'irradiance et de production électrique solaire

Objectifs

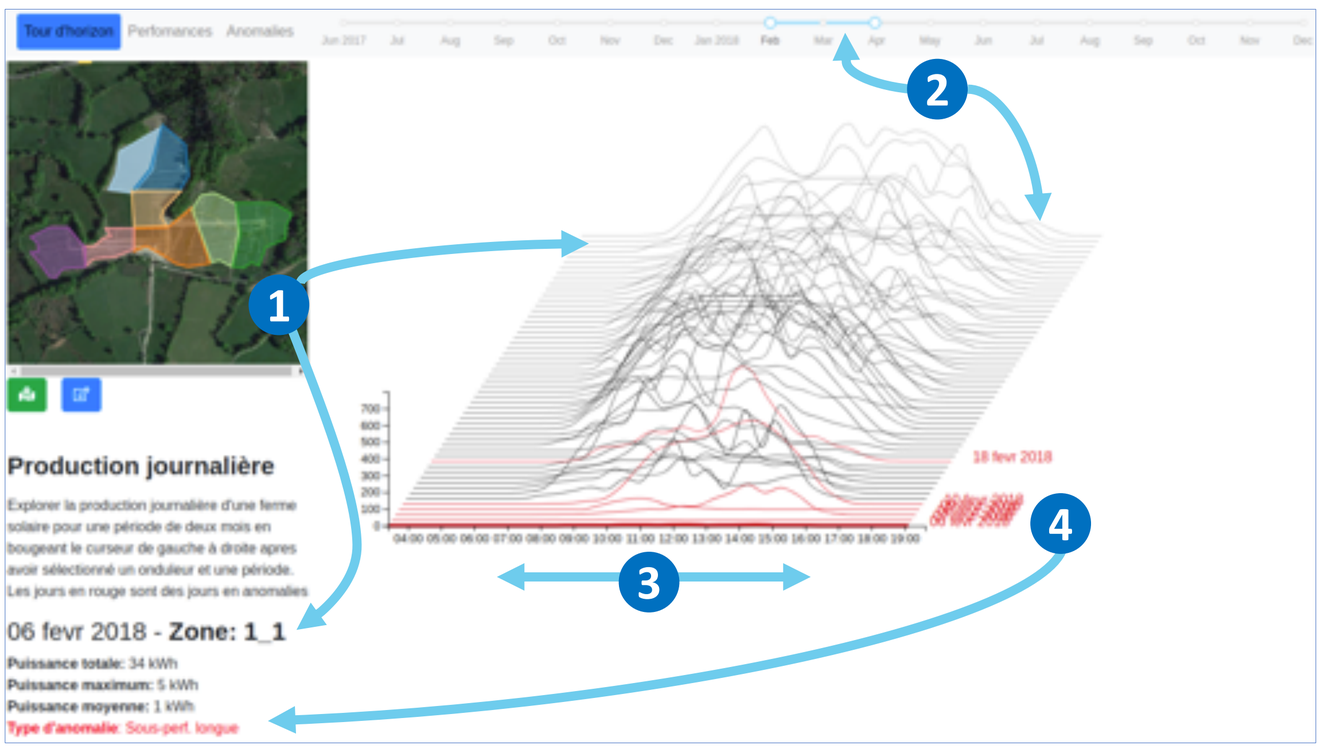
- Créer un outil d'analyse pour les **data scientists**.
- **3 niveaux de visualisation** pour :
  - Faciliter l'analyse de l'historique
  - Évaluer et comparer les sous-performances
  - Évaluer la détection d'anomalies
- **Technologies** :

Auteurs : Alba Ordoñez, Charles Thérond, Ioan Catana, Karine Pétrus, Stéphane Mulard

01 Explorer l'historique de la production

Analyse de l'historique en hiver

- Présence de 5 jours consécutifs d'anomalies (6 au 10 février) pour toutes les zones.
- Corrélation de ces anomalies à la présence de neige sur les panneaux après vérification de la météo.
- Production d'énergie qui augmente graduellement en avançant vers le mois d'avril où les journées sont de plus en plus ensoleillées.



1. **Sélection d'une zone** de la ferme et mise à jour de l'historique de production et de la description sur la zone choisie.
2. **Sélection d'une période de temps de 2 mois** et mise à jour de l'historique de production sur la période choisie.
3. **Bouger le curseur de gauche à droite** permet de naviguer à travers les jours de la période sélectionnée.
4. **Sélection d'un jour d'historique** et mise à jour de la description des caractéristiques de la journée quant à la puissance mesurée et à la présence / type d'anomalie.

02 Comparer les performances des 8 zones

Sous-performances récurrentes

- Ces graphiques ont permis de détecter une période de sous-performance significative pour la zone 2_2 la semaine du 30 août 2018.
- Comme on peut le voir sur le détail de la production du 2 septembre, cette zone affiche une sous-performance tout au long de la journée par rapport aux autres zones, qui ne s'explique pas par un phénomène météorologique.

1. **Sélection d'une ou plusieurs zones** de la ferme et mise à jour de la production globale hebdomadaire, de la production par jour sur une semaine et de la production totale par zone sur la période sélectionnée.
2. **Sélection d'une période de temps** et mise à jour des graphiques de la production hebdomadaire, quotidienne et de la production totale par zone sur la période.
3. **Sélection d'un « point semaine »** et mise à jour de la production hebdomadaire sur la semaine et du détail de la production du premier jour de la semaine.
4. **Sélection d'un « point jour »** mise à jour du détail de la production pour cette journée.

03 Explorer les jours en sous-performance

Limites des méthodes par seuil

- Sur cet exemple, le seuil fixe absolu donne un F1-Score de 0.85, ce qui est bon malgré 4 faux négatifs, avec 2 retards de production les 2 et 22 octobre dont le détail est présenté en bas à droite.
- L'algorithme d'Isolation Forest montre des limites pour la séparation des points sur cet exemple, mais le retard de production du 2 octobre est bien identifié.

1. **Sélection d'une zone** et mise à jour de la somme des résidus, de la distribution d'anomalies et du nuage de points obtenu par PCA des résidus sur la période choisie.
2. **Sélection d'une période de temps** et mise à jour de la somme des résidus de la distribution d'anomalies et du nuage de points obtenu par PCA des résidus.
3. **Sélection de la méthode de seuillage** pour détecter les anomalies. Les jours situés au-dessus du seuil sont les anomalies détectées, les jours avec des pastilles colorées représentent les vraies anomalies.
4. **Sélection de l'algorithme de clustering** et mise à jour du contour de détection des anomalies.
5. **Sélection d'un « point jour »** mise à jour du détail de la production de ce jour.