UNIVERSITÀ DI PISA

Laurea Magistrale in Data Science and Business Informatics

VISUAL ANALYTICS

VAST Challenge 2018

UNIVERSITÀ DI PISA

Autore:

Massidda Fabrizio

Sommario

1 – Introduzione

In questo progetto viene affrontata una Mini-Challenge della *Visual Analytics Science and Technology (VAST) challenge*, una competizione annuale creata con lo scopo di promuovere il campo della visualizzazione e analisi di dati.

La VAST Challenge 2018 presenta tre mini-challenges da risolvere basate su uno scenario fittizio. In questo scenario esiste una città chiamata *Mistford,* situata vicino alla riserva naturale di *Boonsong Lekagul,* nella quale è presente una piccola area industriale dove operano un certo numero di aziende. All'interno della riserva sembra che i nidi di una particolare specie (i *Rose-Crested Blue Pipit*) stiano diminuendo drasticamente. Ciò ha portato ad un'indagine (sulla quale è basata la VAST Challenge 2017) che coinvolge l'azienda *Kasios Office Furniture*, la quale opera nella città di *Mistford*. L'azienda *Kasios* si presenta però come un'organizzazione eco-friendly e nega qualsiasi accusa di scarico di rifiuti tossici nella riserva.

La Mini-Challenge 2 si concentra sull'analisi di dati idrologici raccolti all'interno della riserva, dove i campioni raccolti provengono da diverse zone e contengono misurazioni di diversi agenti chimici di possibile interesse. Più precisamente, la challenge pone tre specifiche domande:

- Caratterizza la situazione passata e più recente circa la contaminazione di sostanze chimiche all'interno della riserva. Durante questo tipo di analisi noti qualche trend potenzialmente interessante?

- Quali anomalie hai trovato nei dati idrologici? Come tali anomalie hanno impattato sulla tua analisi sui problemi ambientali? I dati raccolti sono sufficienti per capire la situazione della riserva? Quali cambiamenti suggeriresti nella raccolta dei dati per rendere l'analisi più efficiente?

- Esistono ragioni per preoccuparsi dell'incolumità dei *Pipit* o di altre specie?

2 – I dati

Sul sito internet della challenge è possibile scaricare un file ZIP contenente i dati idrologici, una breve descrizione su di essi ed inoltre una mappa della riserva sulla quale sono indicate le zone in cui sono stati effettuati i campionamenti e l'area dove si presume siano stati scaricati i rifiuti tossici.

I dati idrologici sono contenuti in un file CSV. Ogni riga del file rappresenta la misurazione di un agente chimico in un determinato luogo e momento. Gli attributi sono 5:

- **Id**: numerico; identificativo di ogni record;
- **Value**: numerico continuo; indica il valore assunto dall'agente chimico in quel record;
- **Location**: stringa; nome del luogo in cui è stata effettuata la misurazione;
- **Sample date**: data; giorno-mese-anno relativo al record;
- **Measure**: stringa; nome dell'agente chimico misurato.

Il numero totale di righe del dataset è 136824, le sostanze chimiche considerate sono 106 e le location 10. I dati sono raccolti in un periodo che va dal 1998 al 2016.

3 – Implementazione

Per lo sviluppo del progetto è stato utilizzato *Vue.js,* un framework JavaScript utile per costruire interfacce utente e *single-page applications*. Per tale ragione è risultata fondamentale la conoscenza di JavaScript, HTML e CSS.

Le principali librerie utilizzate durante il progetto sono state:

- **D3**, o **Data-Driven Documents**, utilizzata per la realizzazione di visualizzazioni interattive e dinamiche nei web browsers;

- **Bootstrap Vue**, che permette di utilizzare bootstrap, una delle più famose librerie al mondo per le componenti front-end;

- **Plotly**, per la realizzazione di grafici.

Durante la fase di preprocessing sono stati inoltre utilizzati il linguaggio di programmazione Python e le librerie Numpy e Pandas, per la manipolazione dei dati e la produzione di nuovi dataset utili all'analisi visuale.

4 – Progetto

Il primo grafico sulla pagina è un lollipop plot, che ha sostanzialmente le caratteristiche di un bar plot, dove ogni barra è rappresentata da una linea e da un cerchio. Esso viene utilizzato per rappresentare la relazione tra una variabile categorica e una numerica. In questo caso troviamo il numero di misurazioni sull'asse delle ascisse e il nome della sostanza chimica sull'asse delle ordinate. Gli agenti chimici sono ordinati per numero totale di misurazioni in ordine decrescente.

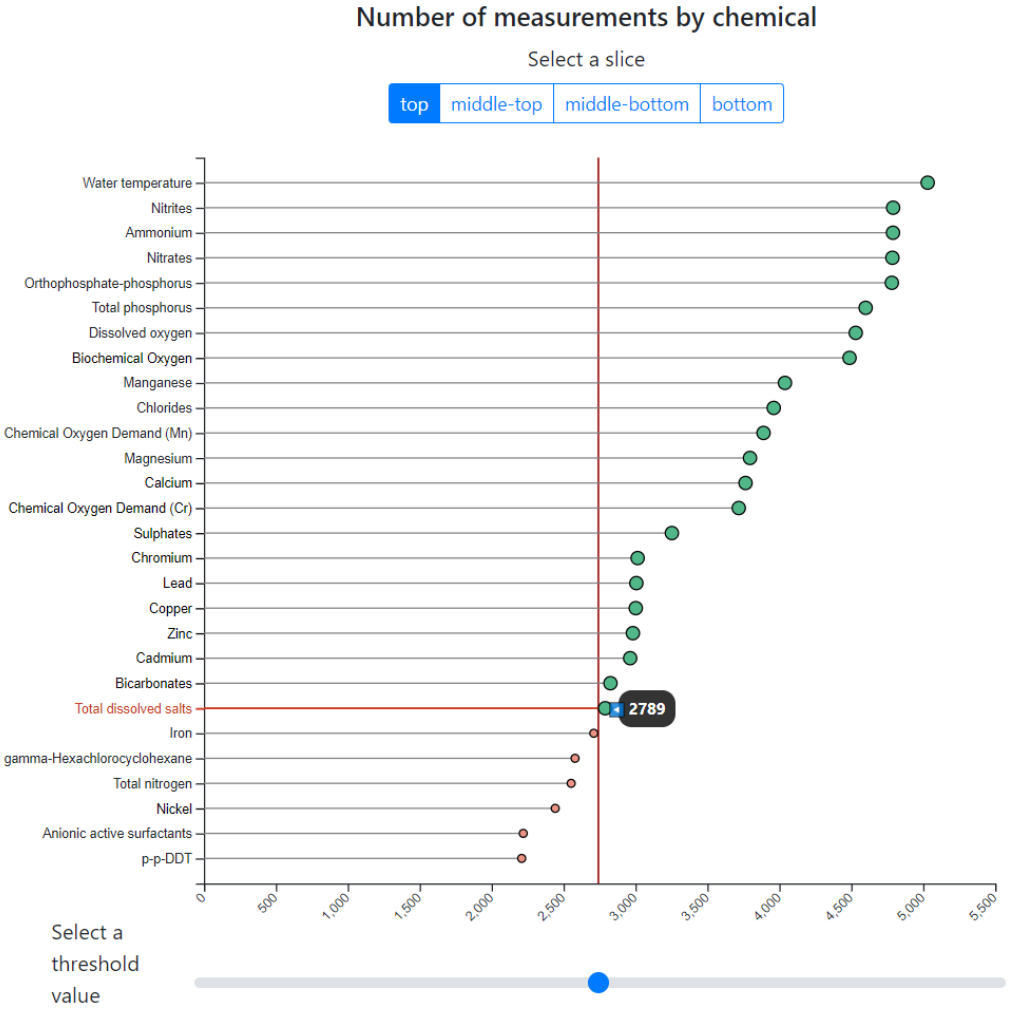

Figura 1 - Lollipop plot

Sopra il grafico sono presenti quattro bottoni che permettono di selezionare uno specifico gruppo di dati. Tale scelta è stata presa in modo da facilitare la visualizzazione delle 106 sostanze, e inoltre perché il plot è collegato a quello sotto, che verrà descritto in seguito. Sotto al grafico è invece presente una barra tramite la quale si può scegliere un valore di threshold per il numero di misurazioni. Le sostanze chimiche il cui numero di misurazioni è inferiore alla threshold saranno escluse dal grafico sotto, e saranno inoltre indicate anche sul lollipop plot attraverso un cambio di colore dei cerchi, da verde a rosso, e da una diminuzione del raggio dei cerchi.

Il secondo grafico è un insieme di bar plots, in cui, ognuno di essi rappresenta il numero di misurazioni per sostanza chimica nel corso degli anni.

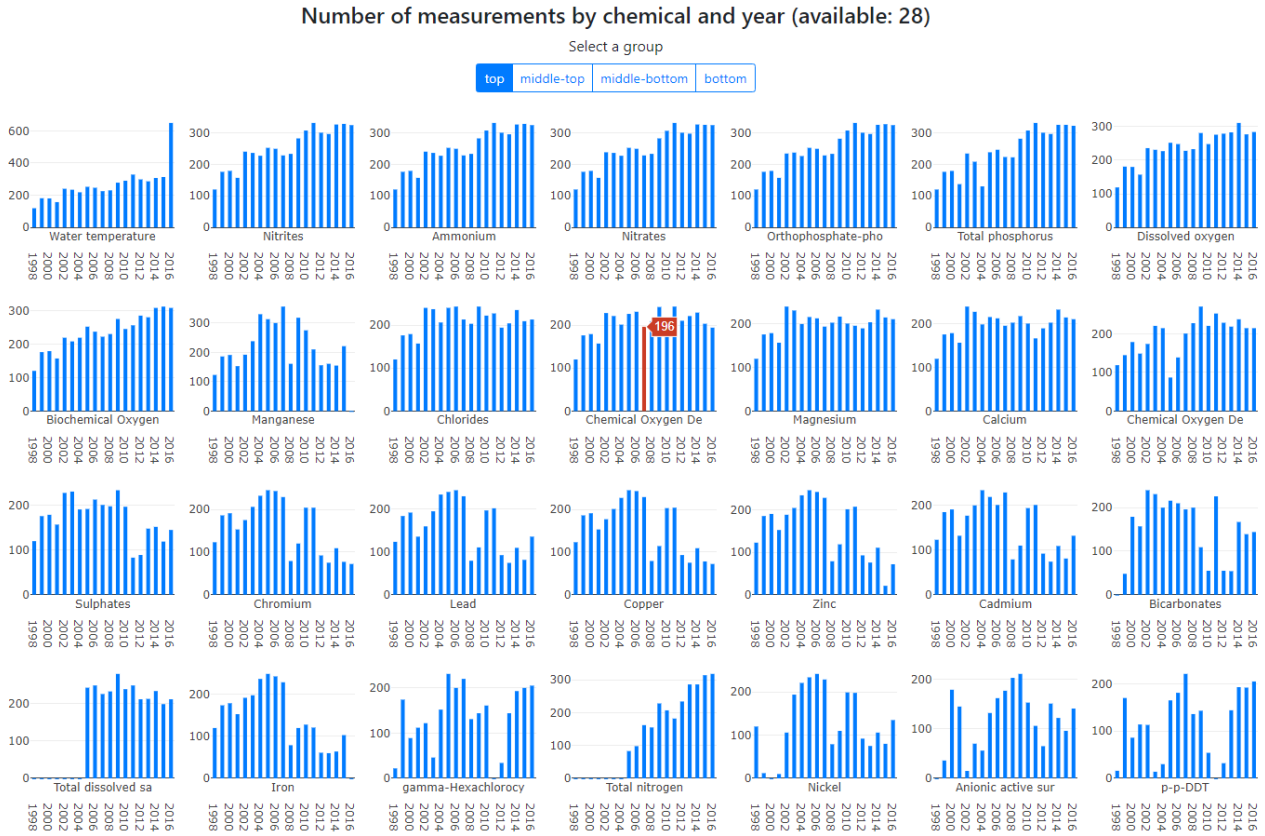

Figura 2 - Bar Subplots

Se la threshold è impostata a 0, anche in questo caso saranno presenti quattro bottoni, ognuno corrispondente ad un gruppo di dati, che coincidono con quelli del lollipop plot sopra. All'aumentare della threshold e del numero di sostanze escluse, il grafico e i bottoni si aggiornano automaticamente. Cliccando sopra al bar plot di una determinata sostanza si aggiunge la stessa al grafico successivo, cliccando nuovamente viene rimossa.

Nel terzo grafico viene rappresentato l'andamento dei valori di una o più sostanze chimiche nel corso degli anni. Per aggiungere una sostanza al plot è possibile cliccare sul grafico sopra, come descritto pocanzi, oppure aggiungere manualmente il nome. Una volta aggiunto, apparirà un tag con il nome relativo e la time series nel grafico. Per rimuovere la sostanza si può utilizzare il tag oppure si può ricorrere nuovamente al grafico sopra. Un'altra alternativa per la rimozione dell'agente chimico è quella di utilizzare il tasto 'Reset All', che compare una volta che nel grafico è presente almeno una time series.

Il limite massimo di time series osservabili contemporaneamente è impostato a 5 per questioni di visibilità.

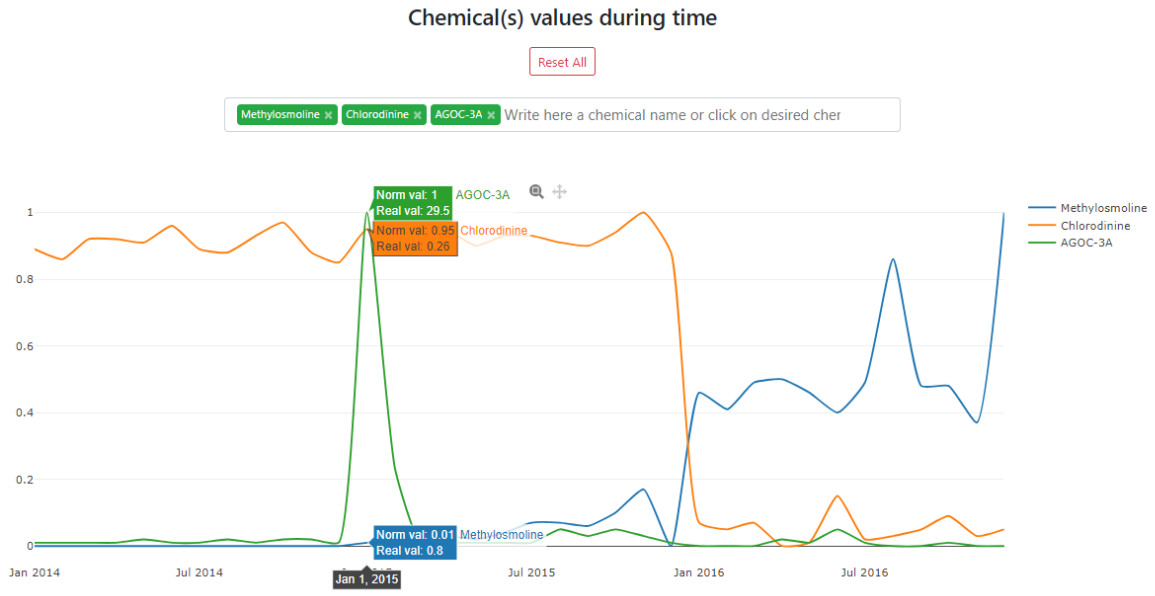

Figura 3 - Time series plot delle sostanze chiave della Challenge 2017

In questo plot è possibile zoomare semplicemente selezionando con il mouse la zona desiderata, mentre per tornare al livello di zoom originale è sufficiente procedere con un doppio click all'interno dell'area del grafico. Sono presenti inoltre i due pulsanti 'Zoom' e 'Pan' (quest'ultimo per muoversi all'interno del grafico nel caso di zoom).

Per permettere il confronto tra diversi agenti chimici, i valori osservati sono stati normalizzati per chemical (min-max normalization), di conseguenza ogni time series avrà valori compresi tra zero e uno. Scorrendo con il mouse sul grafico vengono visualizzati i valori del/dei chemical(s) tramite una tendina che mostra sia il valore normalizzato (Norm), sia quello non standardizzato (Real).

Il quarto e ultimo grafico è una heatmap, una rappresentazione grafica dei dati in cui i valori individuali contenuti in ogni cella della matrice sono rappresentati da colori; in questo caso viene utilizzato per mostrare l'andamento della sostanza chimica selezionata in tutte le locations contemporaneamente.

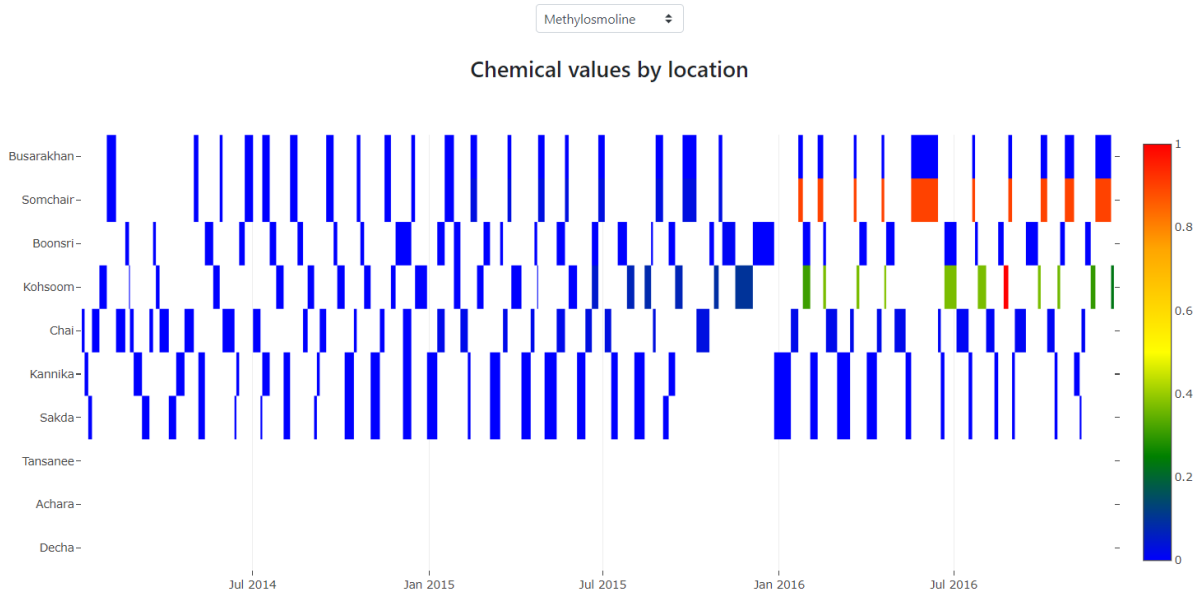

Figura 4 – Heatmap della sostanza Methylosmoline

Anche in questo caso è possibile zoomare e tornare alla visualizzazione originale tramite click e sono disponibili le stesse funzioni descritte in precedenza. Le sostanze chimiche disponibili per questo grafico sono quelle presenti nel time series plot, e per selezionare quella desiderata è sufficiente cercarla nel menù a tendina posto sopra la heatmap.

Scorrendo con il mouse sulle celle non vuote della matrice viene visualizzata la data specifica dell'osservazione e il valore assunto dalla sostanza.

5 – La challenge

Per rispondere alle domande poste dalla challenge sfruttiamo i grafici presentati nel paragrafo precedente.

Dal momento che si vuole descrivere l'andamento nel tempo delle sostanze chimiche, è innanzitutto necessario filtrare le sostanze in qualche modo: esse, come detto in precedenza, sono 106 e sarebbe impossibile visualizzarle tutte.

Osservando il lollipop plot (Figura 1) si nota subito che alcune sostanze chimiche hanno un numero elevato di misurazioni, in particolar modo quelle presenti nel gruppo 'Top', ma tale numero decresce molto velocemente nei restanti gruppi.

Avvalendosi anche del secondo grafico (Figura 2) emerge inoltre che alcune sostanze sono misurate frequentemente in tutti gli anni dal 1998 al 2016, mentre altre hanno un numero di misurazioni insufficiente per descrivere l'andamento temporale. Per tale ragione possono essere escluse dall'analisi tutte le sostanze dei gruppi 'Bottom' e 'Middle-bottom', ad eccezione delle sostanze AGOC-3A, Chlorodinine e Methylosmoline (sostanze chiave nella challenge dell'anno precedente), le quali presentano un buon numero di valori durante gli ultimi 3 anni (2013-2016).

Possono essere inoltre escluse dall'analisi anche la maggior parte delle sostanze del gruppo 'Middle-High', per le stesse ragioni descritte sopra.

Osservando innanzitutto i chemical chiave della VAST Challenge del 2017 (Figura 3), si nota che:

- Non esistono misurazioni per tali chemical nelle località di Tansanee, Achara e Decha;

- L'andamento di AGOC-3A, la quale dovrebbe rappresentare l'alternativa eco-friendly alle altre due sostanze, presenta un unico picco durante il mese di Gennaio 2015 nella località di Boonsri, che potrebbe essere dovuto ad un errore di misurazione. Il trend di tale sostanza è sospetto, in quanto ci si aspetterebbe una crescita nei valori che però non avviene;

- La Chlorodinine invece mostra un trend decrescente a partire da Gennaio 2016 che corrisponde alle aspettative, ad eccezione anche in questo caso di un solo picco durante il mese di Giugno 2016 nella località di Kohsoom.

- La Methylosmoline, sostanza per la quale è stata accusata la ditta Kasios, mostra un rallentamento fino a Dicembre 2015, dopodiché ricomincia a crescere in modo evidente. Osservando l'andamento dei valori nelle varie località (Figura 4), spiccano quelli delle stazioni Somchair e Kohsoom. I valori della prima località risultano strani in quanto impennano improvvisamente e poi rimangono costanti: non si esclude che tale situazione sia dovuta a malfunzionamenti nei sensori. I valori di Kohsoom mostrano invece un trend crescente a partire da metà 2015.

Altri trend potenzialmente interessanti sono:

- La sostanza Nitrites presenta un trend crescente nella località di Tansanee durante la seconda metà del 2014. A dicembre 2014 raggiunge il picco, dopodiché ritorna a valori normali (Figura 5a).

- La sostanza Ammonium cresce a partire dalla metà del 2013 nelle località di Kohsoom e di Tansanee, fino a tornare a valori normali intorno a Gennaio 2015 (Figura 5b);

- La sostanza Magnesium (Figura 5c) presenta valori molto elevati nel periodo Febbraio - Maggio 2011 in tutte le località eccetto Sakda, Achara e Tansanee (in quest'ultima non sono presenti dati);

- La sostanza Anionic active surfactants presenta un trend crescente dal 2014 in poi nella località di Boonsri e, in modo più preoccupante, nella località di Kohsoom (Figura 5d);

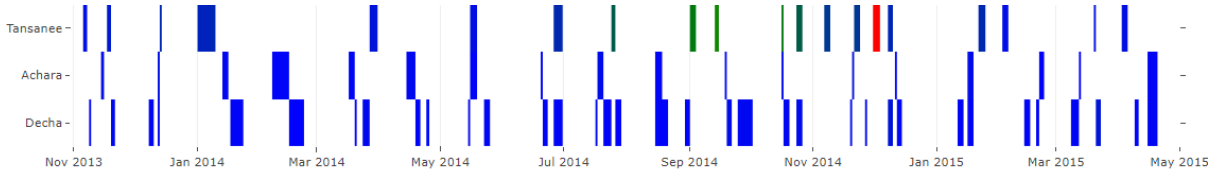

Figura 5a – Nitrites

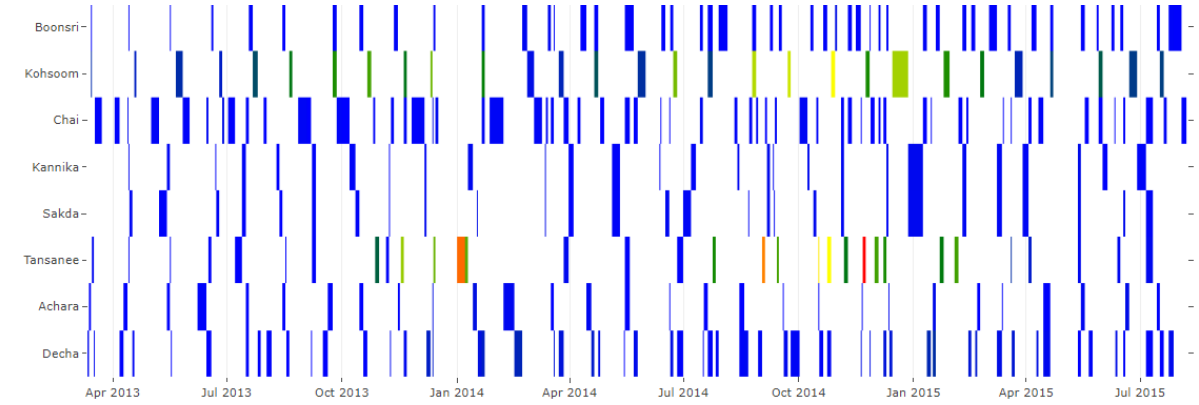

Figura 5b – Ammonium

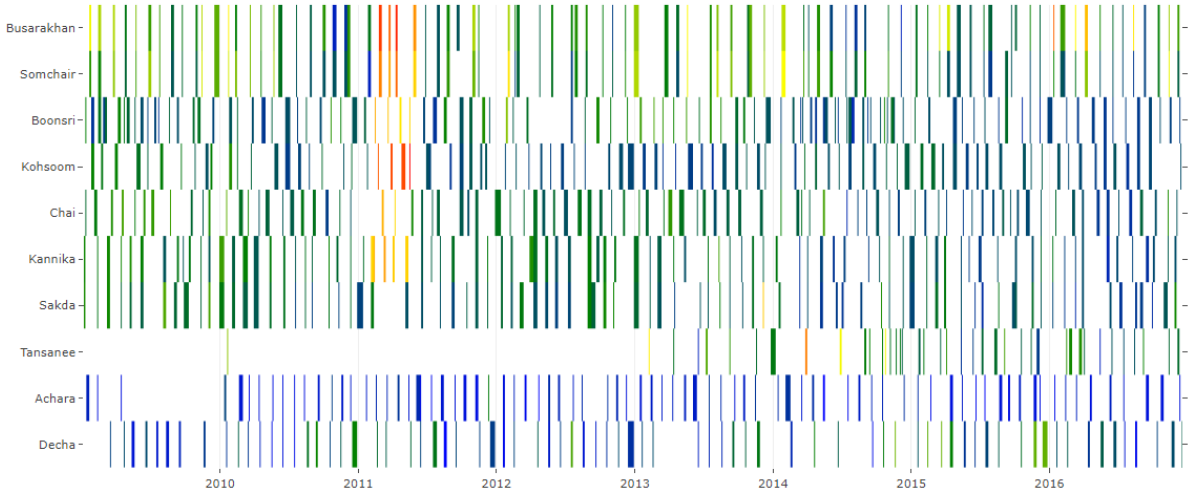

Figura 5c – Magnesium

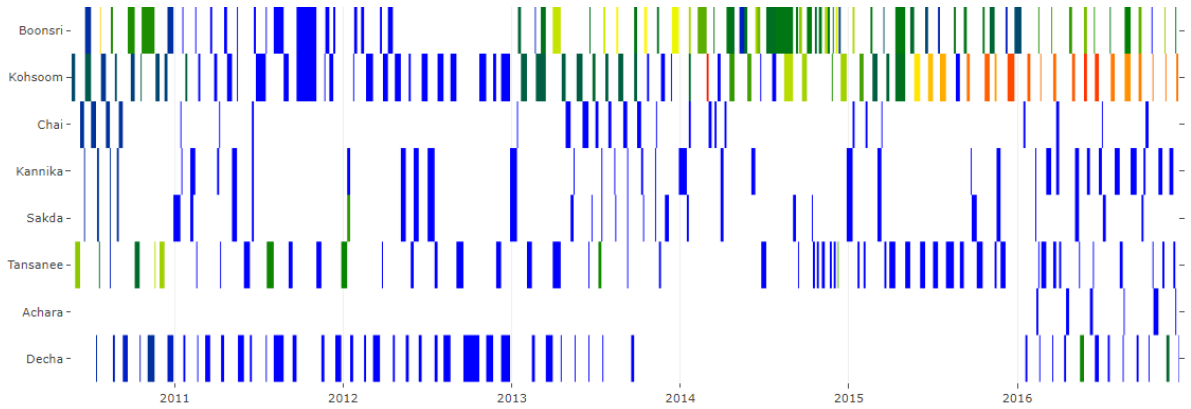

Figura 5d – Anionic active surfactants

Nei dati idrologici è possibile osservare diverse anomalie. Innanzitutto, come già detto precedentemente, il numero di misurazioni è diverso a seconda della sostanza e della località considerata. Ciò significa che le misurazioni non sono effettuate in modo automatizzato, cosa che renderebbe sicuramente più efficiente sia la raccolta dei dati che l'analisi successiva.

Per le località Tansanee, Achara e Decha sono disponibili dati solamente dal 2009 in poi: ciò limita chiaramente l'analisi sulle specifiche località. È limitata anche l'analisi su specifiche sostanze, dal momento che più della metà vengono escluse a causa del numero insufficiente di misurazioni nel tempo.

Altre anomalie osservate:

- La sostanza chimica Atrazine ha mantenuto valori esattamente pari a 0.5 dal 2008 al 2011;

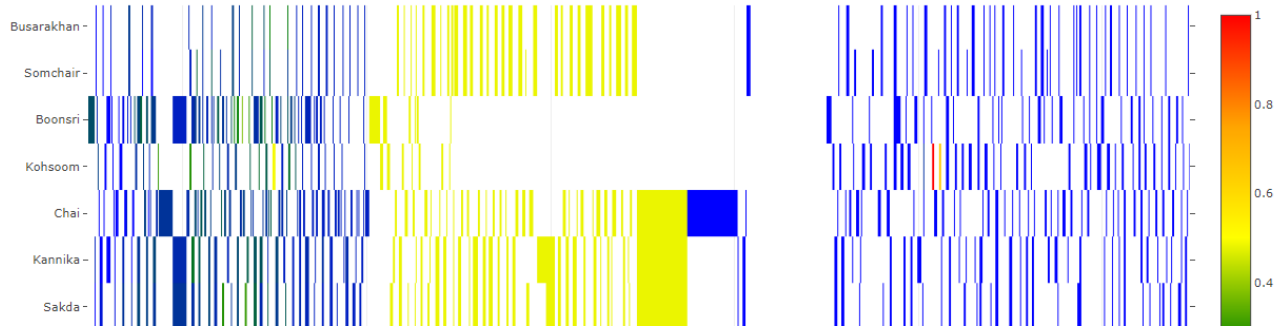

Figura 6 - Atrazine

- Ogni misurazione della sostanza Mercury ad Achara nel periodo Gennaio 2010 – Aprile 2011 ha valore esattamente pari a 1.

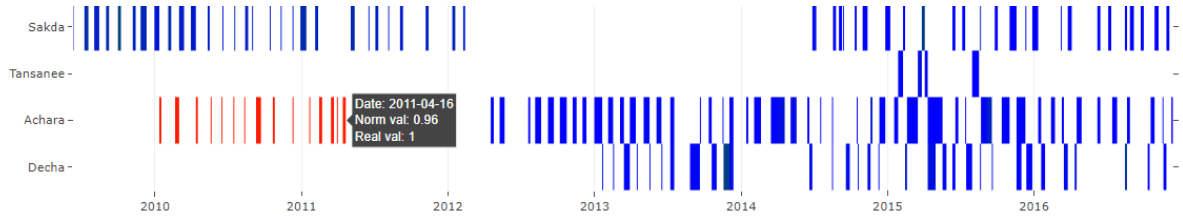

Figura 7 - Mercury

Infine, si è notato che alcune sostanze chimiche evidenziano determinate zone nelle quali i valori sono storicamente più alti:

- Nella località di Tansanee si possono osservare valori elevati nelle seguenti sostanze: Nitrates, Chlorides, Total nitrogen, Arsenic;

- Nella località di Kohsoom: Ammonium, Orthophosphate-phosphorus, Fecal coliforms e Fecal streptococci;

- In entrambe Tansanee e Kohsoom: Chemical oxygen demand(mn), Chemical oxygen demand(cr), Calcium, Sulphates, Bicarbonates, Total dissolved salts;

- Nella località di Decha: Chlorides e Zinc.