Visual Analytics 2019/2020 Report
VAST Challenge 2018 - MC2 Like a duck to water

Virginia Morini, matricola: 522134

April 2020

Abstract

Report del progetto realizzato per il corso di Visual Analytics 2019/2020 presso Università di Pisa. Il progetto nasce con l'obiettivo di rispondere, tramite elementi visuali ed interattivi, alle domande poste nella mini challenge 2 della VAST Challenge 2018. La prima sezione presenta la traccia proposta dalla VAST Challenge; La seconda descrive la fase di data understanding; Nella terza sezione sono descritti i widget visuali e interattivi utilizzati all'interno dell'apllicazione; Infine la quarta sezione è dedicata alla discussione e analisi dei risultati.

1 Mini challenge 2 - Like a duck to water

La VAST Challenge è un concorso annuale nato con l'obiettivo di far avanzare il campo dell'analisi visiva attraverso la competizione. Ogni anno la VAST challenge fornisce agli utenti un problema realistico da risolvere suddiviso in tre mini challenge, ognuna avente il proprio dataset.

Per capire a fondo il problema affrontato nella VAST Challenge 2018 è necessario citare la traccia dell'anno precedente.

1.1 VAST Challenge 2017 - Mini Challenge 1 e 3

Mistford è una città di medie dimensioni, situata a sud-ovest di una grande riserva naturale, la Boonsong Lekagul Wildlife Preserve. La città ha una piccola area industriale con quattro stabilimenti principali.

Mitch Vogel è uno studente post-dottorato che studia ornitologia al Mistford College e ha scoperto segni che il numero di coppie che nidificano del Rose-Crested Blue Pipit, un popolare uccello locale, sta diminuendo. Tale diminuzione è sufficientemente significativa da indurre la Pangera Ornitology Conservation Society a sponsorizzare Mitch a intraprendere ulteriori studi per individuare le possibili ragioni.

A tal fine, la Challenge si è concentrata sull' analisi del traffico all'interno della riserva e sullo studio di immagini satellitari degli ultimi anni.

1.2 VAST Challenge 2018 - Mini Challenge 2

L'anno scorso la Kasios Furniture Company, una delle aziende situate a Mistford, è stata coinvolta in danni ambientali alla Boonsong Lekagul Wildlife Preserve sia per lo scarico di rifiuti tossici che per l'inquinamento dell'aria con sostanze chimiche provenienti dal suo processo di produzione. Kasios ha però negato qualsiasi accusa di scarico di rifiuti industriali, affermando di aver sostituito il Methylosmoline (sostanza tossica) con AGOC-3A, sostanza ecologica. I suoi portavoce hanno dichiarato che non c'è alcuna contaminazione del terreno vicino alla stazione dei ranger, come suggerito dai partecipanti alla mini challenge 1 e 3 dell'anno scorso. Hanno, inoltre, ispezionato quell'area, trovandola incontaminata come il resto della riserva.

Dopo tali affermazioni, i professori di ornitologia del Mistford College sono andati a dare un'occhiata alla discarica, eseguendo analisi del terreno. I ranger della Boonsong Preserve hanno però ottenuto risultati inconcludenti nel rilevamento del Methylosmoline in quanto, a causa della costruzione di una nuova stazione dei ranger, nel sito erano stati attuati scavi e attività edilizie.

Con la scomparsa di una prova primaria contro Kasios, gli investigatori dovranno adottare un altro approccio. I professori del Dipartimento di Idrologia del Mistford College hanno fornito diversi anni di letture di sensori d'acqua provenienti da fiumi e torrenti della riserva. Questi campioni sono stati prelevati da diverse località sparse nell'area e contengono misurazioni di diverse sostanze chimiche di possibile interesse, ma non sono mai stati analizzati per mancanza di fondi. L'obiettivo della challenge è quindi quello di indagare i dati idrologici di tutta la riserva rispondendo alle seguenti domande:

1. Caratterizzare la situazione passata e recente per quanto riguarda la contaminazione chimica nei corsi d'acqua del Boonsong Lekagul. Vede qualche tendenza di possibile interesse in questa indagine?

2. Quali sono le anomalie riscontrate nel dataset di misurazioni fornito? In che modo queste influiscono sulla vostra analisi dei potenziali problemi per l'ambiente? Il Dipartimento di Idrologia sta raccogliendo dati sufficienti per comprendere la situazione complessiva della Riserva? Quali modifiche proponete di apportare all'approccio di campionamento per comprendere al meglio la situazione?

3. Dopo aver esaminato i dati, c'è qualcosa che suscita particolare preoccupazione per il Pipit o per altri animali selvatici?

2 Data understanding

Il Dataset fornito per svolgere l'indagine proposta dalla Challenge è composto da 136824 records. Ogni record rappresenta una misurazione effettuata all'interno della riserva ed è descritta da 5 attributi, illustrati nella tabella 1.
Inoltre, insieme ai dati è possibile scaricare una mappa in cui sono segnalati i luoghi in cui sono state eseguite delle misurazioni e la zona in cui è situata la discarica.

Le misurazioni, effettuate in un periodo che va da gennaio 1998 a dicembre 2016, riguardano 10 luoghi differenti (siti vicini a corsi d'acqua) e 106 sostanze chimiche. Esplorando il dataset si nota che non tutte le sostanze chimiche sono state campionate in tutte le date ed i luoghi considerati.

Table 1: Descrizione degli attributi del dataset.

Attributo	**descrizione**	**tipo**
id	Identificatore numerico della misurazione	intero
value	Valore misurato	float
location	Luogo in cui è avvenuta la misurazione	stringa
sample date	Data in cui è avvenuta la misurazione (giorno-mese-anno)	stringa
measure	Sostanza chimica misurata	stringa

Per realizzare gli elementi visuali e interattivi presenti nel progetto, i dati sono stati manipolati utilizzando principalmente i linguaggi di programmazione Python e Javascript:

1. *Python:* è stato utilizzato, grazie all'uso di librerie quali Pandas e Numpy, per realizzare diverse aggregazioni dei dati e per normalizzarli.

2. *Javascript:* è stato utilizzato, insieme alla libreria Crossfilter, per la manipolazione dei dati.

3 Widget visuali e interattivi

In questa sezione sono descritte le tecniche utilizzate per l'implementazione di widget visuali e interattivi e analizzati singolarmente questi ultimi.

3.1 Tecniche utilizzate

Come ambiente di sviluppo del progetto si è utilizzato *Node.js*, un framework che permette di realizzare applicazioni lato server in Javascript, e *NPM*, un gestore di pacchetti che permette di includere, rimuovere e aggiornare le librerie.

L'interfaccia utente è stata realizzata grazie all'uso di *Vue.js*, un framework javascript di tipo progressivo che si basa sul rendering dichiarativo e sulla creazione di componenti.

Il progetto è quindi strutturato in componenti: la componente principale (*App.vue*) viene utilizzata principalmente per creare la struttura dell'applicazione web, per gestire gli elementi interattivi e per caricare e manipolare i dati da passare a tutte le altre componenti annidate in cui sono implementati i grafici. In particolare, lo scheletro dell'applicazione è stato costruito tramite l'uso congiunto dei framework *VueBootstrap* e *MDBootstrap*.

Infine per realizzare visualizzazioni interattive dei dati sono state utilizzate le librerie Javascript *Plotly.js* e *D3.js*.

3.2 Circular bar plot

La prima visualizzazione nasce con l'obiettivo di mostrare, per ogni sostanza chimica, il numero di misurazioni effettuate durante l'intero periodo di tempo. Si è scelta tale aggregazione dei dati per fornire all'utente una visione d'insieme del problema prima di addendrarsi nell'indagine vera e propria.
Per caratterizzare l'andamento della contaminazione chimica all'interno della riserva nel corso degli anni, è infatti fondamentale capire che non tutte le sostanze chimiche hanno lo stesso numero di

misurazioni e che alcune di esse ne hanno così poche da rendere difficoltosa un' analisi di questo tipo.

A tal fine, è stato implementato con D3 un circular barplot, ossia un barplot in cui ogni barra, rappresentante una sostanza, è mostrata intorno a un cerchio, piuttosto che a una linea. Rispetto ad un barplot, questo tipo di grafico in figura 1, permette un uso migliore dello spazio che ben si adatta all'elevato numero di sostanze chimiche (106) da mostrare.

Tramite il radio button è possibile visualizzare quattro porzioni dei dati, ciascuna rappresentata da un circular bar plot: tutte le sostanze chimiche (opzione di default), quelle aventi un numero di misurazioni minore di 200, quelle nel range [200-2000] e quelle maggiori di 2000. Inoltre, come mostrato in figura 2 , passando il mouse sopra una barra compare un tooltip in cui viene indicato il nome della sostanza e il numero di misurazioni.

Infine, i circular barplot giocano un ruolo centrale all'interno del progetto in quanto cliccando su una barra e quindi su una particolare sostanza chimica il resto dell'applicazione si modificherà di conseguenza, permettendo all'utente di visualizzare informazioni relative a quella sostanza.

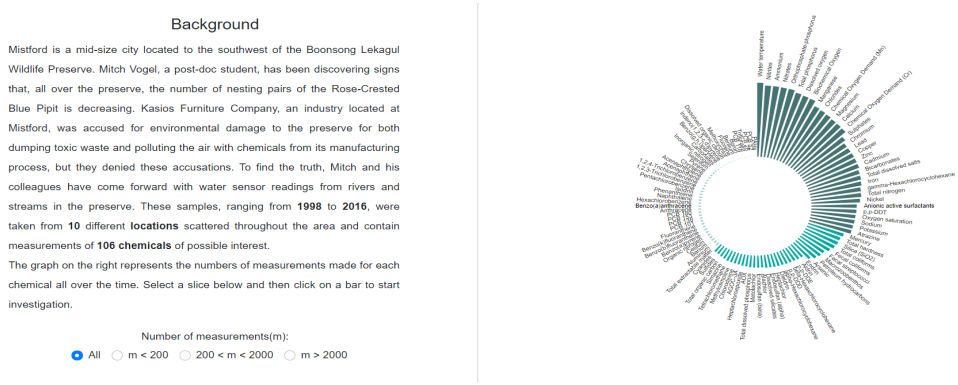

Figure 1: Sulla sinistra introduzione e radio button, sulla destra circular bar plot con tutte le sostanze chimiche.

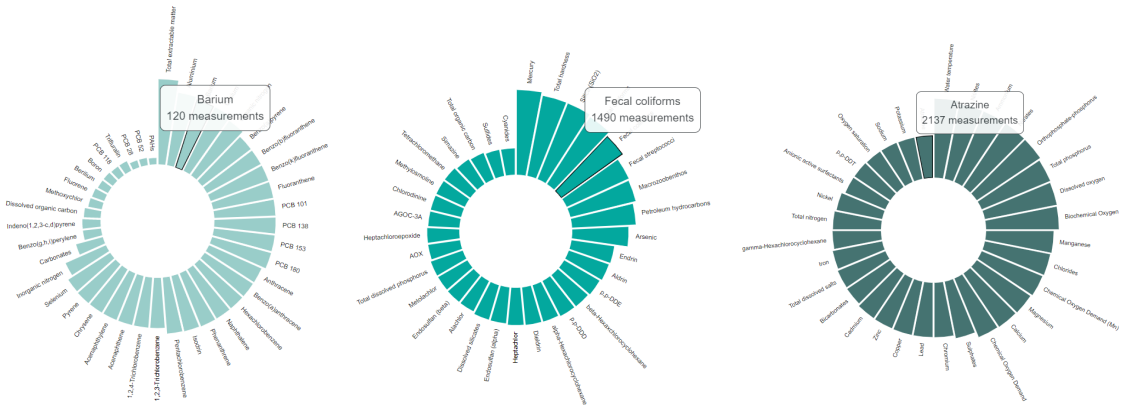

Figure 2: Circular bar plot rappresentanti ognuno una porzione delle sostanze chimiche in base al loro numero di misurazioni.

3.3 Range slider temporale

La restante parte di applicazione è dedicata all'analisi della sostanza chimica selezionata. In alto è stato fissato un range slider temporale che permette all'utente di selezionare il range di anni in cui esplorare i dati (figura 3). I grafici successivi si modificano in base al range selezionato.

Lo slider temporale è composto solamente dagli anni in cui sono state effettuate misurazioni per due motivi: dare all'utente un'idea immediata sulla dimensione temporale effettiva di ogni sostanza chimica; evitare gap nelle visualizzazioni ad esso connesse.

Figure 3: Range slider temporale che permette di esplorare i dati in base agli anni selezionati.

3.4 Line chart

Con l'obiettivo di verificare se sono presenti valori anomali o sospetti, per ogni sostanza chimica è stato implementato in D3 un line chart rappresentante l'andamento dei valori misurati in ogni mese degli anni in cui sono state effettuate misurazioni. Inoltre, nel line chart è presente un' altra linea, parallela all'asse temporale, rappresentante il valore medio dell'intero periodo considerato (figura 4).

Il valore di ogni mese è stato ottenuto calcolando la media di tutti i valori misurati in quel mese, in ogni luogo. La scelta di questa tipologia di grafico è dovuta principalmente al fatto che lavoriamo con valori continui.

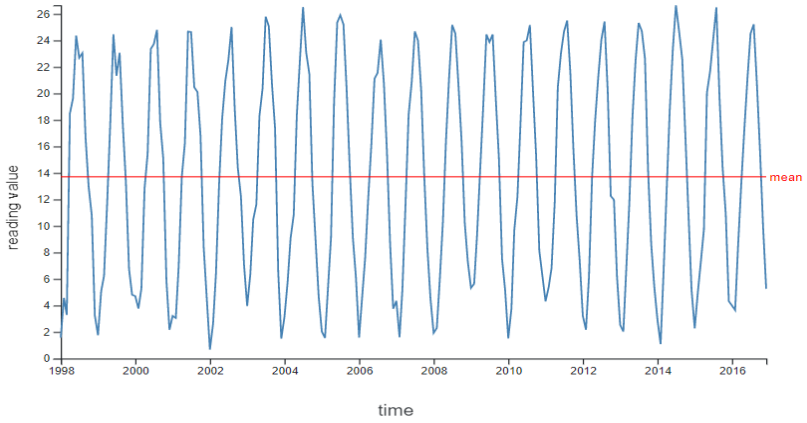

Figure 4: Line chart rappresentante l'andamento dei valori misurati in ogni mese in cui sono state prese misurazioni.

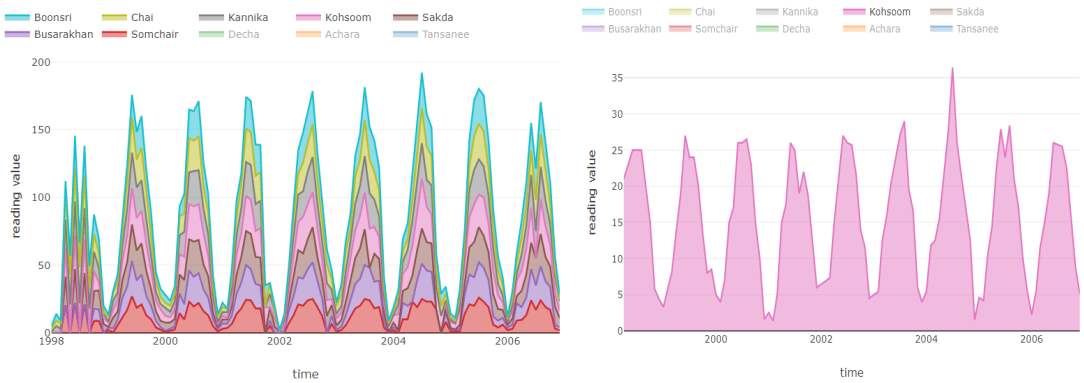

Figure 5: a) Stacked area chart rappresentante l'andamento dei valori in tutte le zone. b) Area chart rappresentante l'andamento dei valori di una singola zona selezionata.

3.5 Stacked Area chart

Per l'indagine, risulta interessante studiare l'andamento dei valori mensili di una sostanza chimica non solo nella sua totalità ma anche suddiviso per zona. Per tale motivo è stato implementato con Plotly uno Stacked Area Chart. Questa tipologia di grafico risulta appropriata per studiare e

confrontare tra loro le proporzioni di ogni zona. Inoltre, cliccando su un elemento della legenda è possibile eliminare quella zona dal confronto; mentre, facendo doppio click è possibile analizzare l'andamento dei valori nella singola zona (figura 5b).

Ai fini dell'analisi può risultare utile anche sapere in quali zone non sono mai state effettuate misurazioni. Per questo, come mostrato in figura 5a tali zone sono presenti nella legenda ma opacizzate.

3.6 Bar chart

Per avere una piena comprensione dei grafici sopra descritti è utile studiare l'andamento nei mesi del numero di misurazioni effettuate. Per esempio, se un valore che risulta essere un outlier è stato misurato un'unica volta potrebbe trattarsi di un errore di misurazione.

A tal fine, è stato implementato con plotly un Bar chart che si adatta a dati discreti (figura 6). Passando il mouse sopra una barra compare un tooltip in cui viene indicato il numero di misurazioni e il mese di riferimento.

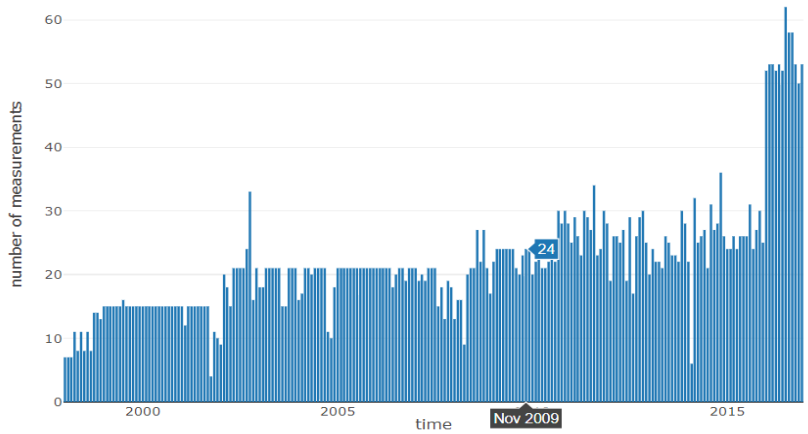

Figure 6: Bar chart rappresentante l'andamento del numero di misurazioni nei mesi.

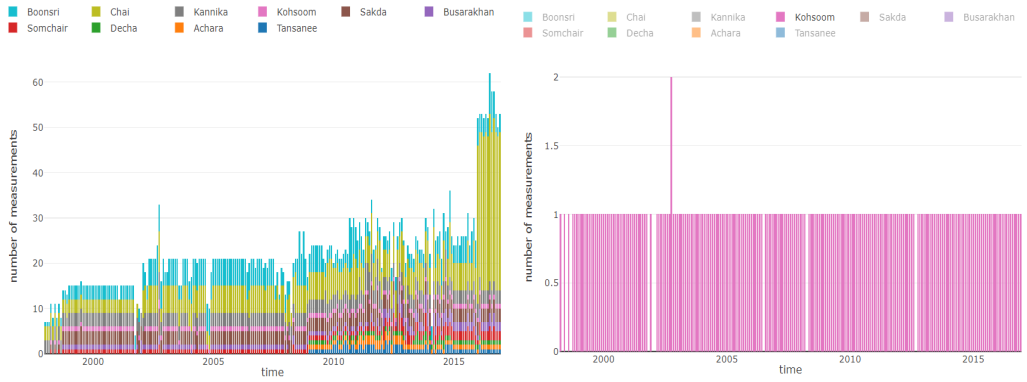

Figure 7: a) Stacked Bar chart rappresentante l'andamento del numero di misurazioni in tutte le zone. b) Bar chart rappresentante l'andamento del numero di misurazioni di una singola zona selezionata.

3.7 Stacked Bar chart

Anche in questo caso, si è deciso di studiare l'andamento del numero di misurazioni non sono nella sua totalità ma per zona. Tramite l'uso di plotly è stato utilizzato uno stacked bar chart, in cui ogni porzione della barra rappresenta una zona (figura 7a).

Cliccando su un elemento della legenda è possibile rimuovere quella zona dal confronto. Facendo doppio click è possibile studiare l'andamento dei valori nella singola zona selezionata. Passando il mouse su una barra compare un tooltip in cui viene indicato il numero di misurazioni per ogni zona e il mese di riferimento (figura 7b). Le zone in cui non sono state effettuate misurazioni sono opacizzate.

3.8 Multi-line chart

L'ultima visualizzazione presente sulla pagina nasce con lo scopo di permettere all'utente di confrontare i valori delle sostanze chimiche tra loro per evidenziarne eventuali anomalie e similarità.

Si è scelto di implementare con d3 un multi-line chart in cui ogni linea rappresenta l'andamento nei mesi dei valori della sostanza chimica selezionata (figura 9). I valori di ogni sostanza sono stati normalizzati, scalandoli in un range [0,1] (min-max scaling), per facilitarne un confronto diretto.

Ogni volta che nel circular bar plot viene selezionata una nuova sostanza chimica, nel multi-line plot viene mostrato solamente l'andamento di tale sostanza. Sopra il grafico è presente una select, contenente tutte le 106 sostanze chimiche ordinate per numero di misurazioni decrescente: selezionandone una o più di una, il grafico si modificherà mostrando l'andamento delle sostanze selezionate e di quella scelta nel circular bar plot (la quale non può essere deselezionata) nel range temporale selezionato nello slider. Inoltre, cliccando su un elemento della legenda, è possibile deselezionare una sostanza chimica che verrà quindi eliminata dal grafico.

Se, nella select, viene selezionata una sostanza già presente comparirà un pop-up che esplicita tale informazione. Se viene selezionata una sostanza che non ha misurazioni nel range temporale scelto comparirà un pop-up che invita l'utente a scegliere, se possibile, un range diverso.

4 Analisi dei risultati

Questa sezione è dedicata alla discussione delle domande poste dalla VAST Challenge 2018.

4.1 Quesito 1

Caratterizzare la situazione passata e recente per quanto riguarda la contaminazione chimica nei corsi d'acqua del Boonsong Lekagul. Vede qualche tendenza di possibile interesse in questa indagine?

Dovendo analizzare 106 sostanze chimiche, misurate in 10 luoghi durante 20 anni, risulta utile capire quali di queste sostanze hanno abbastanza misurazioni nel corso degli anni da rendere efficace un'analisi della situazione recente e passata. Per questo, in base al numero di misurazioni effettuate, le sostanze sono state divise in tre gruppi rappresentati visualmente da circular bar plots.

Guardando il circular in plot in figura 1 risulta evidente la disparità del primo gruppo nel numero di misurazioni rispetto agli altri due. Tale gruppo è costuito da 41 elementi aventi meno di 200 misurazioni totali. Cliccando sui singoli elementi che lo compongono e osservando lo slider temporale (esempio in figura 8), notiamo anche che la gran parte delle sostanze del primo gruppo non raggiunge i 4 anni di misurazioni (precedenti al 2009). Negli altri due gruppi invece, costituiti da 65 elementi, solo il 10% delle sostanze chimiche ha meno di 4 anni di misurazioni.

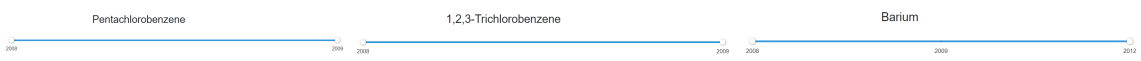

Figure 8: Range slider temporali di sostanze appartententi al primo gruppo e aventi meno di 4 anni di misurazioni.

In base a tali considerazioni, risulta più utile concentrarsi sulle restanti sostanze chimiche per rispondere al quesito. In un primo momento, si è deciso di concentrare l'attenzione sulle sostanze chimiche citate nella traccia della VAST Challenge. Kasios ha infatti affermato di aver sostituito l'uso di sostanze chimiche tossiche, quali Methylosmoline e Chlorodinine, con la sostanza ecologica AGOC-3A. Comparandole in figura 9, notiamo che sono state misurate solamente negli ultimi 3 anni con un ugual numero di misurazioni (474).

Inoltre, vediamo che tutte e tre le sostanze non hanno misurazioni a Decha, Tansanee e Achara, zone piuttosto lontane dalla discarica. Di seguito, sono analizzate in dettaglio:

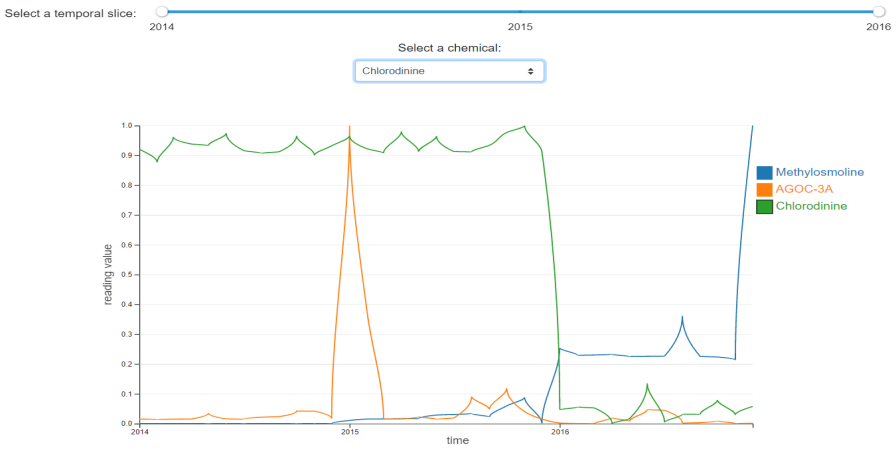

Figure 9: Multi-line chart rappresentante, per ogni sostanza chimica selezionata, l'andamento dei valori nei mesi degli anni selezionati tramite il range slider.

- *AGOC-3A:* Come mostrato in figura 10a, l'andamento dei valori negli anni risulta piuttosto stabile, nonostante un picco deciso nel gennaio 2016 a Boonsri dove il valore misurato è di 16, ben lontano dal valore medio (circa 2). Dallo Stacked Bar plot si può però apprendere che sono state effettuate solamente 2 misurazioni a gennaio in quella zona. Potrebbe quindi trattarsi di un outlier o di un errore di misurazione.
- *Methylosmoline:* Prima del 2016 i valori sembrano essere piuttosto bassi ma a inizio 2016 crescono decisamente a Koshoom e Somchair (figura 10b). Studiando i valori di quest'ultima zona, notando che non c'è varianza nei valori dal gennaio al novembre 2016, possiamo dedurre che si tratti di un errore. L'aumento a Koshoom, appare invece molto più realistico e rilevante dal punto di vista dell'analisi: Kasios sostiene infatti di averne sospeso l'uso nell'ultimo anno.
- *Chlorodinine:* L'andamento dei valori presenta un notevole decremento fino a inizio 2016, ma nella stessa estate si nota un nuovo aumento a Kohsoom, sito più vicino alla discarica (figura 10c). Risulta interessante notare che sia Chlorodinine che Methylosmoline hanno un incremento dei valori a Khosoom nello stesso periodo.

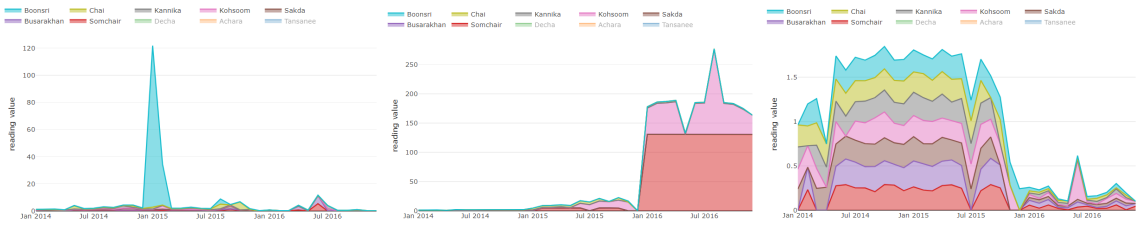

Figure 10: Stacked Area charts rappresentanti andamento dei valori per zona. a)AGOC-3A, b)Methylosmoline, c)Chlorodinine

Inoltre, essendo Kohsoom la zona più vicina alla discarica, si è deciso di studiare l'andamento dei valori in questa zona per ogni sostanza chimica. Oltre alle già citate Methylosmoline e Chlorodinine, le sostanze che presentano trend di interesse sono Ammonium e Total dissolved phosphorus. In entrambe (figura 11a,b), durante l'intero periodo di tempo considerato, i valori registrati a Kohsoom sono decisamente più alti di quelli delle altre zone. Da evidenziare anche Anionic active surfants che dal 2013 in avanti presenta un notevole innalzamento dei valori sia a Kohsoom che a Boonsri.

Infine, notiamo che le sostanze total coliforms, fecal coliforms e fecal streptcocci hanno lo stesso andamento nell'intero periodo, probabilmente perchè si tratta di batteri appartenenti alla stessa famiglia (figura 11c).

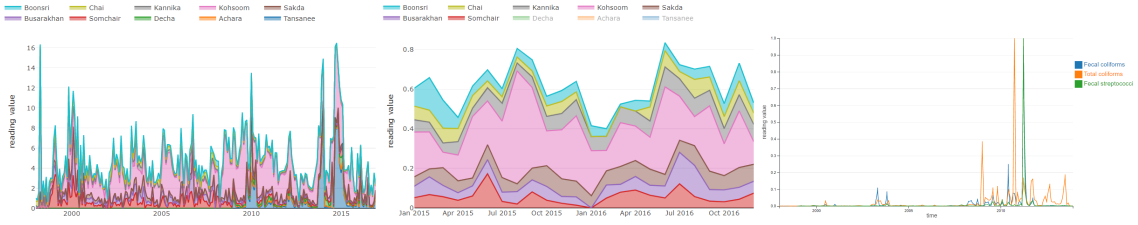

Figure 11: a)Stacked Area chart rappresentante andamento dei valori per zona di Ammonium, b)Stacked Area chart rappresentante andamento dei valori per zona di Total dissolved phosphorus, c)Confronto tra i tre coliforms.

4.2 Quesito 2

Quali sono le anomalie riscontrate nel dataset di misurazioni fornito? Inche modo queste influiscono sulla vostra analisi dei potenziali problemi perl'ambiente? Il Dipartimento di Idrologia sta raccogliendo dati sufficienti per comprendere la situazione complessiva della Riserva? Quali modifiche proponete di apportare all'approccio di campionamento per comprendere al meglio la situazione?

Alcune sostanze chimiche presentano picchi o decrementi notevoli nei valori in un periodo così ristretto di tempo da far pensare a errori di misurazione:

- *Magnesium:* Forte picco in tutte le zone nell'estate 2011. Non si osserva però un comportamento anomalo nel numero di misurazioni effettuate.
- *Chemical Oxygen Demand (Cr):* Decremento sia dei valori che del numero di misurazioni in tutte le zone nel 2015 (figura 12a).
- *Iron:* Forte picco nell'agosto 2003 in tutte le zone. Non si osserva però un comportamento anomalo nel numero di misurazioni effettuate (figura 12b).
- *Atrazine:* Andamento costante (varianza 0) dei valori dal 2008 al 2010, con un piccolo decremento causato dalla mancanza di misurazioni (figura 12c).

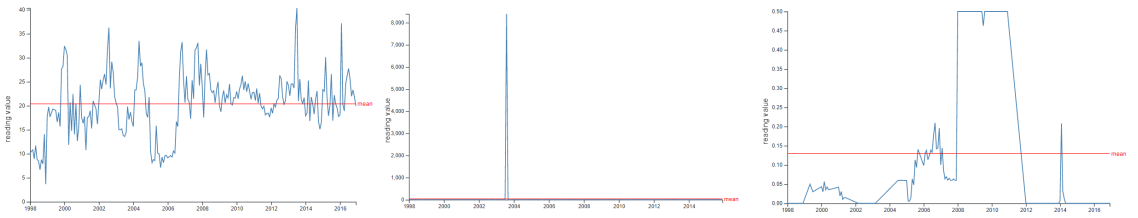

Figure 12: Line charts rappresentanti andamento dei valori e discostamento dal valore medio. a) Chemical Oxygen Demand (cr), b)Iron, c)Atrazine.

Tali valori contribuiscono negativamente sull'analisi in quanto non permettono di avere una visione realistica dei valori nel tempo. Inoltre è complesso capire se si tratta di errori o di outlier.

Un altro fattore che limita le possibilità di analisi sono sicuramente i valori mancanti sia per gli anni che per le zone. Per avere una visione complessiva del livello di contaminazione della riserva, è infatti preferibile avere per tutte le sostanze chimiche lo stesso numero di misurazioni in tutti gli anni e in tutte le zone prese in considerazione, per permettere di effettuare confronti diretti.

L'inserimento di sensori nelle zone d'interesse potrebbe facilitare il processo di raccolta dati.

4.3 Quesito 3

Dopo aver esaminato i dati, c''e qualcosa che suscita particolare preoccupazione per il Pipit o per altri animali selvatici?

Come evidenziato nella risposta al primo quesito, l'andamento dei valori delle tre sostanze centrali AGOC-3A, Methylosmoline e Chlorodinine dimostrano che le affermazioni di Kasios non sono veritiere. L'uso di AGOC-3A, sostituito ecologico del Methylosmoline, infatti dovrebbe incrementare nel tempo, mentre, escludendo un picco a Boonsri probabilmente dovuto ad un errore, rimane piuttosto costante. La Methylosmoline, i cui valori dovrebbe decrescere, presenta invece un incremento deciso a inizio 2016 a Somchair e Kohsoom. Lo stesso picco, meno deciso, viene registrato anche per la Chlorodinine a Kohsoom.

Tutto ciò porta a mantenere alto il livello di preocuppazione per la fauna della riserva.