Comme nous l’avons vu dans la partie précédente, aucune analyse statistique de concerts et de salles qui les accueillent portant sur l’ensemble de la Suisse et prenant en compte les dimensions spatiales, temporelles et artistiques n’a à ce jour été menée. Grâce aux études passées en revue, nous avons remarqué que des outils et des techniques pour étudier les concerts sous ces trois aspects existaient. Nous avons donc constitué une base de données en mettant en relation des données issues de trois services différents : Songkick, Spotify et Wikidata. Nous verrons dans cette partie comment les données ont été récupérées et traitées à partir de ces trois sources, quelles sont les méthodes d’analyse qui ont été développées et enfin comment justifier ces choix méthodologiques.

**\section{Récupération et traitement des données}**

La structure et la mise à disposition des données diffèrent grandement entre Songkick, Spotify et Wikidata. Un procédé propre à chaque service de récupération et de traitement des données a donc été appliqué.

**\subsection{Songkick}**

Afin de pouvoir analyser les concerts qui ont lieu en Suisse, il est essentiel d’obtenir des données sur ces événements. La plateforme Songkick \footnote{\url{www.songkick.com}}, créée en 2007, recense des concerts qui ont lieu partout dans le monde dans le but de mettre en relation spectateurs, artistes et services de vente de billets. Le site revendique six millions de concerts dans sa base de données et quinze millions d’utilisateurs. Pour rassembler toutes ces données, Songkick combine des techniques de \emph{web crawling} \footnote{Le \emph{web crawling}, ou \emph{robot d’indexation} en français, consiste à naviguer automatiquement sur des pages Web pour les indexer et en extraire certaines données.} à partir de plusieurs sources et des partenariats avec des sociétés de vente de billets \autocite{vanbuskirk\_songkick\_2009}. Chaque concert référencé sur Songkick est défini par un lieu et au moins un artiste. Il en ressort une vaste base de données qui comprend l’historique des concerts en fonction des lieux et en fonction des artistes.

**\subsubsection\*{Récupération des données}**

Bien que le site propose une API, la technique de \emph{web scraping} \footnote{Le \emph{web scraping} consiste à extraire des données de certaines pages Web.} s’est révélée plus pertinente, en raison de l’impossibilité via l’API d’obtenir un historique des concerts en fonction du lieu. La première étape de la récupération des données a été d’identifier les salles et lieux de concert se trouvant en Suisse, en utilisant le moteur de recherche de Songkick. Ensuite, il a été possible de concevoir une liste de tous les lieux de concert définis par Songkick comme étant en Suisse. A partir de cette liste de lieux, une liste de tous les concerts qui y ont été organisés a pu être établie. Enfin, une liste de tous les artistes ayant participé à ces concerts a été obtenue. Il en résulte donc trois jeux de données distincts, avec les propriétés suivantes :

\begin{itemize}

\item Concerts : nom, lieu, artiste(s), date.\\

82'011 individus.

\item Lieux : nom, code postal (optionnel), rue (optionnel), localité, pays, coordonnées géographiques (latitude et longitude, optionnel).\\

7'657 individus.

\item Artistes : nom.\\

32'231 individus.

\end{itemize}

**\subsubsection\*{Traitement des données}**

La base de données de Songkick étant enrichie automatiquement, plusieurs erreurs se retrouvent dans les données récupérées, dont des doublons de lieux (des individus distincts qui se réfèrent au même lieu). Ces doublons ont été identifiés par leur nom, leur adresse, leurs coordonnées géographiques ou manuellement. Les doublons ont été réunis sous un unique élément, choisi en fonction des propriétés les plus complètes. 433 doublons ont été recensés, amenant le nombre de lieux distincts à 7’224. Certains lieux n’avaient pas de coordonnées géographiques, uniquement une localité. Une opération de géocodage \footnote{Le géocodage consiste à assigner des coordonnées géographiques à une adresse physique.} a donc été effectuée sur ces 2'326 lieux de Songkick distincts répartis dans 144 localités uniques en utilisant l’API de la plateforme Mapbox \footnote{\url{https://www.mapbox.com}}. Une rapide vérification des résultats obtenus par Mapbox a montré que les coordonnées attribuées aux localités étaient correctes. Les lieux avec des coordonnées géocodées perdent certes en précision, puisque les coordonnées renseignent la localité et pas le lieu exact, mais ils gagnent à être inclus dans le reste des données pour toute analyse spatiale. Plusieurs lieux comportaient aussi des coordonnées géographiques qui se situaient hors de Suisse. Les coordonnées pour ces six lieux ont été remplacées par des coordonnées géocodées en fonction de la localité donnée. Enfin, un individu du jeu de données \emph{artistes} dénommé \emph{Unknown artist} a été supprimé, portant le nombre d’artistes distincts à 32’230.

**\subsection{Spotify}**

Etant donné que Songkick ne permet d’obtenir que le nom de chaque artiste, davantage de données relatives aux artistes sont nécessaires. Le service de streaming musical Spotify a été créé en 2008 et compte dans son catalogue plus de 70 millions de morceaux \footnote{\url{www.spotify.com}}. Devenue l’un des poids lourds du secteur, la plateforme doit une partie de son succès à son système de recommandations, qui repose sur l’agrégation d’une multitude de données relatives à chaque morceau de son catalogue \autocite[46]{prey\_knowing\_2018}. La classification des morceaux se fait de plusieurs manières, notamment en labélisant les genres musicaux des artistes avec précision, comme \emph{Zurich indie}, et en analysant le signal audio des musiques pour quantifier la dimension acoustiques.

**\subsubsection\*{Récupération des données}**

Spotify met à disposition gratuitement une API qui permet d’explorer une partie des données dont elle dispose. La première étape a été de trouver une correspondance entre les artistes sur Songkick et ceux présents sur Spotify. Pour cela, les noms des artistes de Songkick ont été soumis au moteur de recherche de Spotify via l’API. Le moteur de recherche donne plusieurs résultats. Le résultat dont le nom partageait la plus grande similarité \footnote{La similarité entre deux chaînes de caractères a été obtenue avec la fonction \emph{ratio} du module Python \emph{SequenceMatcher}. Cette fonction repose sur l’algorithme du Gestalt Pattern Matching : \url{https://en.wikipedia.org/wiki/Gestalt\_Pattern\_Matching}.} avec le nom cherché a été retenu. Sur les 32'230 artistes de Songkick, une correspondance a été établie avec 28'571 artistes de Spotify, ce qui laisse 3'659 artistes de Songkick sans correspondance. Pour chaque artiste, ses genres musicaux (le cas échéant), son nombre d’auditeurs sur Spotify sur les 30 derniers jours, le nombre de personnes qui le suivent sur Spotify et ses morceaux les plus populaires (au maximum dix) ont été pris en compte. Cela représente 3'744 genres musicaux distincts et 248'625 morceaux. Pour chaque morceau, les données quantitatives issues de l’analyse du signal audio ont été récupérées. Ces traits musicaux sont les suivants \footnote{\url{https://developer.spotify.com/documentation/web-api/reference}} :

\begin{itemize}

\item Acousticness : détermine si un morceau est plutôt acoustique ou non. De 0 (peu acoustique) à 1 (très acoustique).

\item Danceability : détermine si un morceau est adapté pour danser. De 0 (peu dansant) à 1 (très dansant).

\item Duration : durée du morceau (en millisecondes).

\item Energy : détermine si un morceau est intense. De 0 (peu intense) à 1 (très intense).

\item Instrumentalness : détermine si un morceau est instrumental. De 0 (aucun instrument) à 1 (aucune parole). Les valeurs en dessus de 0.5 déterminent en général des morceaux instrumentaux.

\item Liveness : détermine si du public est audible sur un morceau. De 0 (pas de public) à 1 (enregistrement en public).

\item Key : détermine la tonalité d’un morceau.

\item Loudness : détermine le volume sonore global d’un morceau en décibels (dB). En général de -60 à 0 dB.

\item Mode : détermine le mode d’un morceau. 0 pour le mode mineur, 1 pour le mode majeur.

\item Speechiness : détermine si un morceau contient des mots parlés. De 0 (musical) à 1 (poésie).

\item Tempo : détermine le tempo global d’un morceau en battements par minute (BPM). De 0 à 1000.

\item Time signature : détermine la signature rythmique d’un morceau. De 3 à 7.

\item Valence : détermine si un morceau renvoie un sentiment positif. De 0 (négatif) à 1 (positif).

\end{itemize}

L’attribut \emph{liveness}, qui rend compte des caractéristiques de l’enregistrement d’un morceau plutôt que des caractéristiques d’un artiste, et les attributs catégoriels \emph{key}, \emph{mode} et \emph{time signature} ont été écartés. Neuf attributs musicaux pour les 248’625 morceaux ont donc été retenus.

**\subsubsection\*{Traitement des données}**

Comme pour les lieux de concert, des doublons ont été repérés pour les artistes. Plusieurs artistes distincts sur Songkick ont été associés au même artiste sur Spotify. Pour la majorité des cas, cela est dû au référencement automatique d’artistes dans la base de données de Songkick, qui identifie un même artiste plusieurs fois avec des noms différents. Par exemple, le groupe lausannois Larytta a deux identités sur Songkick : Larytta et LARYTTA. Ces doublons ont été réunis sous une unique identité dont le nom est celui qui partage la plus haute similarité avec le nom de l’artiste sur Spotify. Ainsi, 465 doublons ont été traités, pour un total de 28'106 artistes identifiés sur Spotify.

**\subsection{Wikidata}**

Afin de catégoriser la multitude de genres musicaux obtenus sur Spotify, il est nécessaire de les organiser hiérarchiquement. La base de connaissances libre et collaborative Wikidata contient des données structurées qui forment la colonne vertébrale des projets Wikimédia comme Wikipédia. Chaque élément de cette base de données possède un nom, une description, des alias et des propriétés. La structure de ces données aboutit à une organisation hiérarchique définie par des classes et des sous-classes. Par exemple, l’élément \emph{disco} dans Wikidata \footnote{\url{www.wikidata.org/wiki/Q58339}}, décrit comme un \og genre de musique de danse et une sous-culture qui a émergé dans le milieu de la vie nocturne urbaine américaine des années 1970 \fg , est une sous-classe de \emph{dance music} et une instance de \emph{musical form} et de \emph{music genre} et fait partie de \emph{African-American Music}. Ainsi, il est possible de trouver tous les genres musicaux référencés sur Wikidata, mais également de les hiérarchiser grâce aux sous-classes.

**\subsubsection\*{Récupération des données}**

Wikidata possède un service de requête SPARQL, qui permet d’extraire tout type de données de sa base de connaissances et de les télécharger sous forme de tableau au format CSV \footnote{\url{https ://query.wikidata.org}}. Il a fallu extraire tous les éléments de Wikidata qui étaient une instance de \emph{music genre} et retenir pour chacun de ces éléments son ou ses genres parents. Des alias du genre musical (provenant des bases de données Discogs \footnote{\url{www.discogs.com/}} et Everynoise \footnote{\url{https ://everynoise.com}}) ont aussi été récupérés, lorsqu’ils étaient disponibles, pour améliorer les chances de trouver une correspondance avec les genres de Spotify. La requête soumise était la suivante : \par

\begin{lstlisting}[language=SPARQL]

SELECT ?genre ?genreLabel ?parent\_genre ?parent\_genreLabel

?everynoise\_id ?discogs\_genre\_id ?discogs\_style\_id WHERE {

SERVICE wikibase:label {

bd:serviceParam wikibase:language “,en”.

}

?genre wdt:P31 wd:Q188451.

OPTIONAL { ?genre wdt:P279 ?parent\_genre. }

OPTIONAL { ?genre wdt:P9881 ?everynoise\_id. }

OPTIONAL { ?genre wdt :P9218 ?discogs\_genre\_id. }

OPTIONAL { ?genre wdt :P9219 ?discogs\_style\_id. }

}

\end{lstlisting}

Cette requête a renvoyé un tableau avec 6’855 entrées. Cela ne correspond pas à 6’855 genres musicaux distincts puisqu’un genre a plus d’une entrée s’il a plus d’un genre parent.

**\subsubsection\*{Traitement des données}**

Pour hiérarchiser les genres musicaux, il est nécessaire de trouver les genres principaux desquels tous les autres genres découlent. Ces genres principaux sont appelés \emph{top genre} :

\begin{mydef}

Soit deux éléments distincts A et B, instances de \emph{music genre}, l’élément B est le top genre de l’élément A, si A est une sous-classe de B et si la super-classe de B n’est pas une instance de \emph{music genre}.

\end{mydef}

Par exemple, l’élément \emph{classical music} a comme top genre l’élément \emph{art music}, puisque la super-classe de l’élément \emph{art music} est l’élément \emph{music} qui n’est plus une instance de \emph{music genre}, mais de \emph{art form}. Le genre \emph{popular music} a été jugé trop vague, puisqu’il est le top genre entre autres de \emph{rock music}, \emph{pop music}, \emph{jazz music} ou encore \emph{blues}. Cet élément a donc été exclu des top genres. Ensuite, un genre peut avoir plusieurs genres parents et donc plusieurs top genres. Par exemple, le genre \emph{punk rap} a comme genres parents \emph{punk rock} et \emph{hip hop music}, ce qui donne comme top genres \emph{rock music} et \emph{hip hop music}. Tous les top genres distincts ont été conservés pour chaque genre. Les genres qui sont également leur propre top genre n’ont pas été retenus comme top genre. Il en ressort une liste de 4’579 genres musicaux distincts répartis dans 80 top genres. \par

Une fois la classification des genres de Wikidata effectuée, il a fallu trouver une correspondance entre chacun des 3’744 genres recensés sur Spotify et un des 4’579 genres de Wikidata. Tout d’abord, chaque genre musical de Spotify a été comparé à chaque nom ou alias (Discogs et Everynoise) de Wikidata. 3’011 correspondances ont ainsi été trouvées. Pour chacun des 733 genres restants, le top genre le plus fréquent associé aux artistes qui partagent ce genre sans correspondance a été retenu. Par exemple si l’on prend la table \ref{ex\_genre}, où le genre X est un genre de Spotify sans correspondance, les artistes liés à ce genre ont été associés aux top genres M (2 fois) et N (1 fois). Le genre X serait donc associé au top genre M. Grâce à cette technique, 626 correspondances ont été trouvées. Pour les 107 genres restants, une association manuelle avec un top genre a été effectuée.

\begin{table}[h]

\centering

\begin{tabular}{|c|c|c|}

\hline

& genre & top genre \\ \hline

artiste A & genre X & top genre M \\ \hline

artiste B & genre X & top genre N \\ \hline

artiste C & genre X & top genre M \\ \hline

artiste D & genre X & \\ \hline

\end{tabular}

\captionsetup{width=.8\textwidth}

\caption{Exemple d’artistes avec un genre issu de Spotify sans correspondance dans les genres issus de Wikidata. Le \emph{genre X} correspond à un genre de Spotify sans correspondance. Les top genres \emph{M} et \emph{N} correspondent à deux top genres de Wikidata.

\label{ex\_genre}}

\end{table}

Les 48 top genres qui réunissent les 3’744 genres de Spotify ont été en partie affinés manuellement pour éviter que des incohérences issues de la hiérarchisation des données dans Wikidata ne se retrouvent dans le jeu de données récupérées. Des genres jugés trop spécifiques comme \emph{opera} ou \emph{psychedelic music}, identifiés comme des top genres, ont été regroupés au sein de top genres plus larges comme \emph{art music} ou \emph{rock music} respectivement. Ainsi, les 3’744 genres de Spotify ont été réunis en 23 top genres distincts.

\begin{figure}[htp]

\centering

\includegraphics[width=0.5\textwidth]{images/methodo\_genres\_distrib.png}

\captionsetup{width=.8\textwidth}

\caption{Distribution des genres de Spotify dans les top genres}

\label{fig:distrib\_top\_genres}

\end{figure}

**\subsection{Résumé des données à disposition}**

Les données de concert disponibles sur Songkick se révèlent moins fiables pour les événements qui ont eu lieu avant 2007, comme nous l’avons vu dans la partie précédente \autocite{arakelyan\_mining\_2018}. De l’autre côté, les données de concert post-2019 souffrent des incertitudes liées à la crise sanitaire mondiale. Entre annulations et reports de concerts, Songkick ne parvient pas à intégrer tous les changements de programmation. Afin de garantir des données pertinentes qui évitent au mieux les biais inhérents de Songkick et l’incertitude entourant les concerts depuis 2020, seules les données concernant les événements qui se sont déroulés entre le 1\textsuperscript{er} janvier 2010 et le 31 décembre 2019, soit une période de dix ans, ont été retenues dans notre base de données. Le tableau \ref{tab\_recap\_data} résume les données à disposition. Le tableau \ref{anx\_tab\_variables} dans les Annexes indique toutes les variables récoltées, classées par type de données.

\begin{table}[h]

\centering

\begin{tabular}{|l|c|c|c|c|c|}

\hline

& Lieux & Concerts & Artistes & Top genres & Genres Spotify \\ \hline

Données récupérées & 7’657 & 82’011 & 32’231 & 80 & 3’744 \\ \hline

Données traitées & 7’224 & 82’011 & 28’106 & 23 & 3’744 \\ \hline

Données 2010-2019 & 5’063 & 62’827 & 24’719 & 23 & 3’649 \\ \hline

\end{tabular}

\captionsetup{width=.8\textwidth}

\caption{Résumé des données récupérées, traitées et à disposition.}

\label{tab\_recap\_data}

\end{table}

**\section{Méthodes d’analyse}**

**\subsection\*{Mobilité des artistes}**

Estimer la mobilité d’un artiste en Suisse est utile pour repérer si un artiste revêt plutôt un intérêt régional ou national. Un score de mobilité est calculé pour chaque artiste en divisant le nombre de lieux distincts dans lequel il a joué par le nombre de concerts joués (équation \ref{eq\_mobility}). On obtient ainsi un score entre 0 et 1, où un score se rapprochant de 0 indique qu’un artiste change très peu de lieu, tandis qu’un score de 1 montre qu’un artiste n’a jamais joué deux concerts au même endroit. \par

La mobilité $m$ est définie par :

\begin{equation} \label{eq\_mobility}

m = \frac{\text{nbre lieux}}{\text{nbre concerts}}

\end{equation}

Ce score de mobilité ne prend pas en compte la distance entre les différents lieux visités. Pour incorporer cette dimension, le centroïde d’un artiste \footnote{Le centroïde d’un ensemble fini de points (ici les lieux de concert d’un artiste) est la position moyenne de ces points.} est obtenu en faisant la moyenne des coordonnées géographiques des salles fréquentées par cet artiste (équation \ref{eq\_centroid}). La distance moyenne en kilomètres entre le centroïde et chacun des lieux visités est ensuite calculée (équation \ref{eq\_distance}). Enfin, un score de mobilité pondérée est obtenu pour chaque artiste en multipliant le score de mobilité par cette distance moyenne, le tout divisé par 175 (équation \ref{eq\_mob\_weight}). La valeur de 175 résulte du fait que la distance rectiligne maximale en Suisse est d’environ 350 kilomètres. La moitié de cette distance, 175, est la distance moyenne maximale entre un ensemble fini de points se trouvant en Suisse et leur centroïde. Le score de mobilité pondérée obtenu se situe entre 0, qui indique qu’aucune distance n’a été parcourue et qu’un seul lieu a été fréquenté, et 1, qui montre que l’individu ne joue jamais deux fois au même endroit et qu’il parcourt le maximum de distance possible entre ses concerts. \par

Le centroïde $\mathbf{C}$ d’un ensemble fini de $k$ points \(\mathbf{x}\_{1}, \mathbf{x}\_{2}, \ldots , \mathbf{x}\_{k}\) dans $\mathbb{R}^{n}$ est défini par :

\begin{equation} \label{eq\_centroid}

\mathbf{C} = \frac{\mathbf{x}\_{1} + \mathbf{x}\_{2} + \ldots + \mathbf{x}\_{k}}{k}

\end{equation}

La distance moyenne $\overline{ d(\mathbf{x}\_{k}, \mathbf{C})}$ entre un centroïde $\mathbf{C}$ et un ensemble fini de $k$ points \(\mathbf{x}\_{1}, \mathbf{x}\_{2}, \ldots , \mathbf{x}\_{k}\) dans $\mathbb{R}^{n}$ est définie par :

\begin{equation} \label{eq\_distance}

\overline{ d(\mathbf{x}\_{k}, \mathbf{C})} = \frac{\sum\_{n=1}^{k} d(\mathbf{x}\_{n}, \mathbf{C})}{k}

\end{equation}

La mobilité pondérée $m\_p$ est définie par :

\begin{equation} \label{eq\_mob\_weight}

m\_p = \frac{\overline{ d(\mathbf{x}\_{k}, \mathbf{C})} \cdot m}{175}

\end{equation}

La mobilité pondérée maximale parmi les 24'719 artistes atteint un score $m\_p$ de 0.82, obtenu par l’artiste Coely qui a donné deux concerts dans deux lieux différents, soit une mobilité $m$ de 1. Le premier des concerts se trouvait à Onex, à l’extrémité sud-ouest de la Suisse, et le deuxième a eu lieu à Saint-Gall, au nord-est du pays, soit une distance de 286 km entre les deux concerts. Les scores de mobilité pondérée de tous les artistes ont été divisés par ce score maximal de 0.82 pour qu’ils se situent dans l’intervalle $[0, 1]$.

**\subsection\*{Activité des artistes}**

Quantifier l’activité d’un artiste en termes de concerts permet de déterminer si sa présence en Suisse s’étend dans le temps ou bien si elle est plutôt concentrée sur une période précise. La fenêtre d’activité d’un artiste correspond aux dates de son premier et dernier concert, entre le 1er janvier 2010 et le 31 décembre 2019. La fréquence de concerts correspond au nombre de concerts d’un artiste divisé par le nombre de jours que comprend la fenêtre d’activité. La fréquence se situe entre 0, si aucun concert n’a été donné dans cette période, et 1, si un concert a été donné tous les jours dans cette période.

\begin{equation\*}

\text{fréquence de concerts} = \frac{\text{nbre concerts}}{\text{nbre jours d'activité}}

\end{equation\*}

Ensuite, une tournée de concerts a été définie comme une période durant laquelle au moins deux concerts ont lieu sans qu’il ne se passe plus de 90 jours entre deux concerts successifs. Cette valeur limite de 90 jours correspond à la durée d’une pause estivale qui pourrait sectionner une tournée en deux. Une pause entre deux concerts au-delà de trois mois met ainsi fin à une tournée. L’intensité d’une tournée est calculée en divisant le nombre moyen de concerts dans une tournée par le nombre moyen de jours dans une tournée. Une intensité de 0 indique que les concerts ont lieu tous les 90 jours, une intensité à partir de 0.03 montre qu’un concert a lieu par mois, une intensité à partir de 0.14 indique qu’un concert a lieu chaque semaine, et une intensité de 1 signale qu’un concert a lieu chaque jour de la tournée.

\begin{equation\*}

\text{intensité moyenne d’une tournée} = \frac{\text{nbre moyen de concerts dans une tournée}}{\text{nbre jours moyen d’une tournée}}

\end{equation\*}

**\section{Justification des choix méthodologiques}**

Les données ont été récupérées à partir de trois sources différentes, dont deux qui sont en partie automatisées. Il est donc important de voir à quel point ces données sont fiables par rapport à ce qu’elles sont censées représenter.

**\subsection{Précision des données de Songkick} \label{****section\_precision\_songkick}**

Tout d’abord une vérification manuelle des concerts et des artistes répertoriés par Songkick a été menée sur un échantillon du jeu de données. La programmation de trois salles distinctes de trois années différentes a été vérifiée: la Mascotte à Zurich en 2010, la Spirale à Fribourg en 2015 et le Romandie à Lausanne en 2019. Ces trois salles ont été choisies en raison de leur diversité géographique, de leur différente programmation et de la disponibilité de l’historique des événements. Pour chaque salle, les concerts, les artistes principaux et les artistes secondaires durant l’année choisie ont été comparés avec ce que Songkick a référencé et avec ce qui a été retenu dans la base de données. Un score de précision et de rappel a été calculé pour chaque élément selon les formules suivantes \footnote{\url{ https://fr.wikipedia.org/wiki/Pr\%C3\%A9cision\_et\_rappel}} :

\begin{equation\*}

\text{précision} = \frac{\text{nbre éléments correctement recensés}}{\text{nbre éléments recensés}}

\end{equation\*}

\begin{equation\*}

\text{rappel} = \frac{\text{nbre éléments correctement recensés}}{\text{nbre éléments programmés}}

\end{equation\*}

Pour la salle de la Mascotte à Zurich en 2010 (cf. table \ref{tab\_mascotte}), au niveau des concerts et des artistes principaux, le rappel de 68\% par rapport à la précision de 56\% pour les concerts montre que les données retenues pour cette salle sont légèrement plus pertinentes que précises. Cela s’explique par le fait que Songkick a référencé 14 événements qui s’apparentent plus à des soirées qu’à des concerts. Les données retenues souffrent en revanche d’une carence au niveau des artistes secondaires, où plus de la moitié des artistes n’apparaissent pas dans le référencement de Songkick. Cela est en partie dû au fait que les artistes secondaires, programmés en première partie, sont parfois peu ou pas mis en avant dans les programmations, ce qui complique le \emph{web crawling} de Songkick.

\begin{table}[h]

\centering

\begin{tabular}{|l|c|c|c|}

\hline

& Concerts & Artistes princ. & Artistes sec. \\ \hline

Programmation & 41 & 41 & 24 \\ \hhline{|=|=|=|=|}

Songkick Positifs & 30 & 30 & 10 \\ \hline

Songkick Négatifs & 22 & 22 & 4 \\ \hline

Base données Positifs & 28 & 28 & 10 \\ \hline

Base données Négatifs & 22 & 22 & 4 \\ \hhline{|=|=|=|=|}

Base données Précision & 0.56 & 0.56 & 0.71 \\ \hline

Base données Rappel & 0.68 & 0.68 & 0.42 \\ \hline

\end{tabular}

\captionsetup{width=.8\textwidth}

\caption{Comparaison entre les concerts programmés à la salle Mascotte à Zurich en 2010 et ceux répertoriés sur Songkick et dans notre base de données. \emph{Positifs} indique les éléments correctement répertoriés et \emph{Négatifs} indique les concerts erronés.}

\label{tab\_mascotte}

\end{table}

Pour la salle de la Spirale à Fribourg en 2015 (cf. table \ref{tab\_spirale}), la précision est de 100\% pour tous les éléments, mais le rappel n’est que de 30\% pour les concerts et artistes principaux et 43\% pour les artistes secondaires. Cela signifie que Songkick a correctement identifié les concerts qui ont eu lieu, mais n’en a pas repéré suffisamment, ce qui peut être expliqué par la taille relativement modeste de la salle et des événements. Il est possible que certains concerts n’étaient pas référencés sur des sites de vente de billets compris dans le répertoire de Songkick. On remarque également que Songkick a correctement recensé les concerts, mais pas tous les artistes principaux puisque pour quatre concerts l’artiste identifié est « Vernissage Dalbum ». Cette erreur n’a pas été transposée dans la base de données, puisqu’aucune correspondance avec un artiste Spotify n’a pu être établie.

\begin{table}[h]

\centering

\begin{tabular}{|l|c|c|c|}

\hline

& Concerts & Artistes princ. & Artistes sec. \\ \hline

Programmation & 53 & 53 & 7 \\ \hhline{|=|=|=|=|}

Songkick Positifs & 22 & 18 & 3 \\ \hline

Songkick Négatifs & 0 & 4 & 0 \\ \hline

Base données Positifs & 16 & 16 & 3 \\ \hline

Base données Négatifs & 0 & 0 & 0 \\ \hhline{|=|=|=|=|}

Base données Précision & 1 & 1 & 1 \\ \hline

Base données Rappel & 0.3 & 0.3 & 0.43 \\ \hline

\end{tabular}

\captionsetup{width=.8\textwidth}

\caption{Comparaison entre les concerts programmés à la salle La Spirale à Fribourg en 2015 et ceux répertoriés sur Songkick et dans notre base de données. \emph{Positifs} indique les éléments correctement répertoriés et \emph{Négatifs} indique les concerts erronés.}

\label{tab\_spirale}

\end{table}

Pour la salle du Romandie à Lausanne en 2019 (cf. table \ref{tab\_romandie}), les scores de précision et de rappel frôlent la perfection. La précision est de 96\% pour tous les éléments, où les 4\% restants sont le fait de deux concerts programmés mais finalement annulés, ce que Songkick n’a pas pris en compte. Le rappel est de 100\% pour les concerts et les artistes principaux, mais que de 58\% pour les artistes secondaires. Comme pour la Mascotte, cela est probablement dû au fait que les premières parties sont moins mises en avant et échappent donc aux robots de Songkick. La plateforme réussit en revanche à parfaitement distinguer les concerts des soirées festives organisées au Romandie.

\begin{table}[h]

\centering

\begin{tabular}{|l|c|c|c|}

\hline

& Concerts & Artistes princ. & Artistes sec. \\ \hline

Programmation & 45 & 45 & 40 \\ \hhline{|=|=|=|=|}

Songkick Positifs & 45 & 45 & 24 \\ \hline

Songkick Négatifs & 4 & 4 & 2 \\ \hline

Base données Positifs & 45 & 45 & 23 \\ \hline

Base données Négatifs & 2 & 2 & 1 \\ \hhline{|=|=|=|=|}

Base données Précision & 0.96 & 0.96 & 0.96 \\ \hline

Base données Rappel & 1 & 1 & 0.58 \\ \hline

\end{tabular}

\captionsetup{width=.8\textwidth}

\caption{Comparaison entre les concerts programmés à la salle du Romandie à Lausanne en 2019 et ceux répertoriés sur Songkick et dans notre base de données. \emph{Positifs} indique les éléments correctement répertoriés et \emph{Négatifs} indique les concerts erronés.}

\label{tab\_romandie}

\end{table}

**\subsection{Précision des correspondances avec Spotify}**

La précision des correspondances entre les artistes sur Songkick et ceux sur Spotify a été évaluée en analysant manuellement deux échantillons des 28'106 artistes de la base de données (cf. table \ref{tab\_matches}). Le premier échantillon correspond aux artistes dont le score de similarité entre le nom issu de Songkick et le nom issu de Spotify est en dessous de 0.7. Sur ces 338 artistes, 226 ont une correspondance avec Spotify qui a été jugée incorrecte, ce qui représente un taux d’erreur de 67\% des artistes avec une basse similarité. Ce résultat plutôt élevé est à mettre en perspective avec le fait que cet échantillon ne représente que 1\% des artistes de la base de données, qui sont de plus ceux dont les noms sont les plus différents entre Songkick et Spotify. Le deuxième échantillon, composé de 300 artistes sélectionnés aléatoirement, présente lui un taux d’erreur de 2\%. Ce résultat est encourageant et permet de supposer que les équivalences établies entre des artistes de Songkick et de Spotify sont dans l’ensemble très fiables.

\begin{table}[h]

\centering

\begin{tabular}{|l|c|c|c|c|}

\hline

& Nombre & Erreur & Taux d'erreur & Proportion BD \\ \hline

Similarité \textless 0.7 & 338 & 226 & 0.67 & 0.01 \\ \hline

Echantillon aléatoire & 300 & 5 & 0.02 & 0.01 \\ \hline

\end{tabular}

\captionsetup{width=.8\textwidth}

\caption{Vérification manuelle des correspondances entre Songkick et Spotify.}

\label{tab\_matches}

\end{table}