Système d'information décisionnel, de la narration à la simulation : application à la surveillance épidémiologique de la tuberculose au Gabon

Raymond Ondzigue Mbenga

UNIVERSITÉ DE TOURS

ÉCOLES DOCTORALES : MIPTIS/EDR

ÉQUIPES de RECHERCHE : LIFAT/DEBIM-UREMCSE

THÈSE présentée par :

Raymond ONDZIGUE MBENGA

soutenue le : **04 avril 2023**

pour obtenir le grade de : **Docteur de l'université de Tours**
Discipline/ Spécialité : INFORMATIQUE / EPIDEMIOLOGIE

Système d'information décisionnel, de la narration à la simulation : application à la surveillance épidémiologique de la tuberculose au Gabon

Membres du Jury :

M. ENGOHANG-NDONG Jean	Professeur, Kent State University, Président
M. BIMONTE Sandro	Directeur de Recherche, IRSTEA, Rapporteur
M. BEDIANG Georges Wylfred	Professeur (MCA- CAMES), Université de Yaoundé I, Rapporteur
M. ALIGON Julien	Maître de Conférences, Université de Toulouse 1/IRIT, Examinateur
M. DEVOGELE Thomas	Professeur des Universités, Université de Tours, Directeur
M. NGOUNGOU Edgard Brice	Professeur (MC- CAMES), Université des Sciences de la Santé, Co-Directeur
Mme. PERALTA Verónika	Maître de Conférences, Université de Tours, Co-Encadrante

Résumé

La tuberculose (TB) demeure un grave problème de santé publique au Gabon. En effet, selon l'Organisation Mondiale de la Santé (OMS), le pays est désormais compté parmi les 30 pays du monde à forte charge tuberculose. Malgré cette situation épidémiologique très inquiétante, les experts et autorités de la santé publique ne disposent pas aujourd'hui d'outils informatiques adaptés pour surveiller cette maladie.

C'est dans ce contexte que nous nous intéressons dans cette thèse à l'élaboration d'un système d'aide à la décision pour le suivi, la surveillance, l'analyse et la simulation de politiques sanitaires de riposte pour aider les autorités sanitaires à lutter efficacement contre cette pandémie.

La construction d'un tel système pose plusieurs problèmes de recherche, notamment (1) d'intégration des données hétérogènes (actuellement traitées manuellement), (2) la résolution de multiples problèmes de qualité, (3) l'analyse de données spatio-temporelles, (4) la restitution des indicateurs à un public d'experts en santé publique, mais pas informatique et (5) la simulation de diverses politiques sanitaires.

Pour répondre à ces enjeux, nous proposons un système décisionnel qui couple un système d'information géographique décisionnel de la tuberculose (SOLAP-TB) avec un système multi-agents de la tuberculose (SMA-TB). L'objectif est de donner aux experts et autorités de la santé publique la possibilité, à travers un même outil, d'une part de surveiller l'évolution spatio-temporelle de cette pandémie via des tableaux de bord interactifs et dynamiques d'indicateurs puis d'autre part de simuler les politiques sanitaires de riposte via un modèle prédictif. En outre, afin de faciliter la compréhension de la situation épidémiologique de la TB aux experts de la santé publique qui sont des décideurs, nous proposons un processus de narration de données en intelligence épidémique (P-N-D-I-E) qui découle des bonnes pratiques en narration de données et en intelligence épidémique de l'OMS. Ensuite, en utilisant ce processus, nous avons produit une narration de données sur la pandémie de la tuberculose au Gabon qui présente les trouvailles extraites des analyses des données réalisées au niveau du SOLAP-TB.

Finalement, le système décisionnel SOLAP-SMA de la TB que nous avons proposé est générique, c'est-à-dire qu'il peut être adapté à la surveillance de la TB dans d'autres pays. Par ailleurs, le processus de narration de données en intelligence épidémique peut être utilisé pour produire des narrations de données sur d'autres maladies.

Mots clés : système d'aide à la décision spatiale (SOLAP), système multi-agents (SMA), processus de narration de données en intelligence épidémique, tuberculose, Gabon.

Dédicaces

Je dédie ce travail à mon père et à ma mère.

Je le dédie particulière à ma femme et à mes enfants Ventura Prestige, Béni-Ange Parfait et Margueret qui ont souffert de mon absence pendant mes longs séjours doctoraux.

Remerciements

Ce travail de recherche n'aurait jamais pu se concrétiser sans les personnes qui, de près ou de loin, m'ont accompagnées tout au long de celui-ci. Je tiens donc à remercier vivement celles et ceux qui, d'une façon ou d'une autre, m'ont permis de le finaliser.

Mes premiers remerciements sont d'abord adressés à mes directeurs Professeur Thomas DEVOGELE, Professeur Edgard Brice NGOUNGOU et Madame Verònika PERALTA (Maître de Conférences) qui m'ont guidés et aidés tout au long de ce parcours académique. Durant ces trois années, j'ai beaucoup appris à vos côtés.

Mes remerciements touchent également mon laboratoire d'accueil en France, le Laboratoire d'Informatique Fondamentale et Appliquée de Tours (LIFAT) et l'École Doctorale Régionale d'infectiologie tropicale de l'Université des Sciences de la Santé (USS) du Gabon.

Je remercie les équipes BDTLN de l'Université de Tours et DEBIM-UREMCSE de l'Université des Sciences de la Santé du Gabon pour m'avoir accueillies dans le cadre de cette thèse en cotutelle.

Professeur Georges Wylfred BEDIANG et Docteur Sandro BIMONTE, en votre qualité de rapporteurs de mon travail, vous avez consacré une grande partie de votre temps à la lecture de cette thèse afin de formuler des remarques et des suggestions, je vous témoigne ma profonde reconnaissance et vous prie de bien vouloir recevoir mes sincères et vifs remerciements.

Professeur Jean ENGOHANG-NDONG et Monsieur Julien ALIGON (Maître de Conférences) merci d'avoir accepté de participer à mon Jury de thèse.

Je remercie les directions du Programme National de Lutte contre la Tuberculose (PNLT) et de l'Hôpital de Spécialisé de Nkembo ainsi que les médecins qui m'ont apportés leurs expertises médicales dans la partie applicative de ma thèse.

Enfin, je remercie la Banque Mondiale à travers le Projet eGabon-SIS du Ministère de la Santé pour le soutien financier qui a permis de mener à bout ce travail de recherche.

Table des matières

Table des figures

Liste des tableaux

Chapitre 1

Introduction Générale

L'informatique joue un rôle important dans la surveillance des maladies et la riposte. En effet, les systèmes informatiques de surveillance permettent, entre autres, l'amélioration de la notification des épidémies, de la qualité des données ainsi que la capacité de suivi, en quasi-temps réel, de la maladie. En outre, avec ces systèmes, les données peuvent être plus facilement stockées et consultées pour l'aide à la décision.

Ce travail de thèse se situe dans cette lignée. Plus particulièrement, il vise à définir et à mettre en place un système d'information décisionnel pour la surveillance de la pandémie de tuberculose.

1.1 Contexte

Le continent africain fait face à une persistance de la pandémie de tuberculose (TB). En effet, selon l'Organisation Mondiale de la Santé (OMS), sur les dix(10) millions de personnes infectées dans le monde en 2019, 25% se trouvaient dans la région africaine (OMS, 2020). Ce qui place cette région au deuxième rang des régions les plus touchées par la tuberculose dans le monde après l'Asie du Sud-Est (44%).

En 2020, parmi les pays les plus infectés par la TB en Afrique et spécifiquement ceux situés en Afrique centrale (Cameroun, République Démocratique du Congo, République Centrafricaine, Congo, Gabon, Guinée Équatoriale, Sao Tomé-et-Principe, Tchad et Angola), le Gabon est le deuxième pays le plus impacté par la maladie avec une incidence [1] de 527 cas pour 100 000 habitants après la République Centrafricaine (540 cas pour 100 000 habitants) (WHO, 2020). Le taux de létalité [2] de la TB au Gabon est aussi élevé (22%) (WHO, 2020). En outre, le nombre de patients sous traitement antituberculeux perdus de vue (PDV) [3] reste tout aussi inquiétant, passant de 19% à 40% de cas entre 2014 et 2018 d'après les résultats de la revue du Programme National de Lutte contre la Tuberculose

1. En épidémiologie, le taux d'incidence rapporte le nombre de nouveaux cas d'une pathologie observés pendant une période donnée - population incidente- à la population dont sont issus les cas (pendant cette même période)- population cible -. Il est un des critères les plus importants pour évaluer la fréquence et la vitesse d'apparition d'une pathologie (https ://www.insee.fr/fr/metadonnees/definition/c1060)

2. Le taux de létalité est la proportion de décès liés à une maladie, par rapport au nombre total de cas atteints par la maladie

3. Selon OMS, un perdu de vue ou PDV est un patient tuberculeux qui n'a pas entamé de traitement ou qui a interrompu celui-ci pendant deux mois consécutifs ou plus.

(PNLT) 2018. Cette situation épidémiologique très alarmante a conduit l'OMS à classer le pays parmi les 30 pays du monde à forte charge tuberculose.

Pour apporter une réponse à ce problème majeur de santé publique, le PNLT, dans sa stratégie nationale de lutte 2021-2025, s'est fixé un ensemble d'objectifs. Ces objectifs s'alignent à la stratégie de l'OMS qui est de mettre fin à la tuberculose d'ici à 2030 en réduisant de 80% le nombre de cas de tuberculose dans le monde. Parmi ces objectifs, nous pouvons citer (i) accroître la détection du nombre de cas de tuberculose notifiés (toutes formes confondues) de 55% à 90% d'ici fin 2025 (OS1), (ii) augmenter le taux de succès thérapeutique des malades tuberculeux pharmaco-sensibles de 50% à 90% d'ici à 2025 (OS2). À ces objectifs, nous pouvons ajouter la réduction du taux de perdus de vue qui reste encore très élevé. Mais, pour espérer atteindre ces objectifs, les autorités de la santé publique doivent disposer d'outils adaptés pour lutter efficacement contre cette maladie.

1.2 Problématique

Malgré toutes les stratégies de lutte antituberculeuse mises en place depuis plusieurs années par les autorités de la santé publique du Gabon à travers le PNLT, le pays peine toujours à freiner la propagation de cette maladie infectieuse au sein de la population. Cette situation est liée en partie à la faiblesse du système de surveillance de la maladie (PNLT, 2014) : collecte, analyse et transmission des données de la tuberculose à tous les niveaux de la pyramide sanitaire. En effet, les données sont actuellement collectées avec des outils papiers (registres des cas de tuberculose, dossiers médicaux des patients tuberculeux, etc.) puis traitées et analysées manuellement. Cette méthode ne permet pas au PNLT de disposer des données de surveillance de très bonne qualité, ce qui biaise les indicateurs. En outre, elle ne rend pas facile l'intégration de ces données hétérogènes ainsi que leur accès. Ce qui ne permet pas un calcul facile et rapide des indicateurs de surveillance et la production automatique des rapports d'activités pour la prise de décisions stratégiques. En d'autres termes, les acteurs de l'actuel système de surveillance de la tuberculose ne disposent d'aucun outil pour surveiller et suivre en quasi-temps réel l'évolution des indicateurs, prédire la situation future de la maladie ou évaluer les politiques sanitaires de lutte contre la tuberculose.

C'est dans ce contexte que nous nous intéressons dans cette thèse à l'élaboration d'un système d'aide à la décision pour le suivi, la surveillance, l'analyse et la simulation de politiques sanitaires pour aider les experts en santé publique qui sont des décideurs à lutter efficacement contre cette maladie.

La construction d'un tel système nécessite de lever un certain nombre de verrous scientifiques, notamment :

1.2.1 Résoudre de nombreux problèmes de qualité de données hétérogènes de la tuberculose ainsi que leur intégration

Le système de surveillance épidémiologique de la tuberculose actuel est non informatisé. En effet, la collecte, le prétraitement et l'analyse des données des patients tuberculeux se font manuellement en utilisant les outils papiers (par exemple : registre des cas de tuberculose et dossiers médicaux des patients tuberculeux, rapports d'activités) à tous les niveaux de la pyramide sanitaire. Cette méthode est source de nombreux problèmes de qualité de données (par exemple : complétude, fraîcheur, validité) et rend difficile le processus d'intégration de ces données hétérogènes pour la production de rapports d'activités en temps voulu.

1.2.2 Améliorer les moyens de communication utilisés pour attirer l'attention des décideurs sur la situation épidémiologique de la tuberculose

Actuellement, les moyens de communication utilisés par le programme national de lutte contre la tuberculose sont constitués d'affiches et flyers pour sensibiliser les populations sur la tuberculose (par exemple : mode de transmission, symptômes). L'utilisation d'autres moyens plus modernes comme la narration de données peuvent aider à sensibiliser plus facilement les autorités sanitaires, mais aussi le grand public sur la situation épidémiologique de la tuberculose. Mais cette narration de données doit tenir compte des spécificités du domaine de l'intelligence épidémique.

1.2.3 Concevoir un modèle de système multi-agents de la tuberculose réaliste

Pour étudier la propagation d'une maladie comme la tuberculose, les modèles les plus utilisés sont les modèles basiques en épidémiologie « Sains-Infectés-Rétablis » et « Sains-Exposés-Infectés-Rétablis ». Ces modèles simples ne tiennent pas compte de l'ensemble des états réels d'un individu dans le cycle de transmission et traitement de la maladie. Aussi, les données utilisées pour simuler ces modèles ne sont pas souvent extraites d'un entrepôt de données spatiales pour générer les populations de ces modèles. Enfin, ces modèles ne permettent pas de simuler les politiques sanitaires de lutte antituberculeuse. Cependant, ils permettent de connaître l'évolution au cours du temps des populations d'individus selon les états.

1.3 Objectifs de la thèse

1.3.1 Objectif général

L'objectif général de ce travail de recherche est d'améliorer le système de surveillance de la tuberculose du Gabon en proposant un système d'aide à la décision pour le suivi, la surveillance et la simulation des stratégies de lutte antituberculeuse.

1.3.2 Objectifs spécifiques

— Concevoir et déployer un système d'aide à la décision spatio-temporel (SOLAP) de la tuberculose pour aider les décideurs de la santé publique à suivre et à surveiller à quasi-temps réel l'évolution des indicateurs spatio-temporels de cette maladie.
— Proposer une méthode permettant de communiquer facilement et efficacement sur les problèmes de santé publique, destinées aux experts et aux décideurs.
— Concevoir un modèle de système multi-agents de la tuberculose complet et permettant de simuler des stratégies de lutte antituberculeuse avec des données réelles des patients tuberculeux stockées dans l'entrepôt de données d'un système d'information géographique décisionnel (SOLAP).

1.4 Cadre de de la thèse

Ce travail de thèse a été réalisé dans un pays d'Afrique centrale, le Gabon. L'étude a concerné la région sanitaire Libreville-Owendo-Akanda (LBV-OWE-AKA) (Figure 1). Cette région a une population estimée à 817 777 habitants (DGS, 2013), soit 45,15% de la population gabonaise qui est estimée à 1 811 079 habitants (DGS, 2013). Elle est subdivisée en trois communes : (i) Libreville, (ii) Owendo et (iii) Akanda qui constituent l'agglomération de la capitale du Gabon, Libreville. La commune de Libreville comprend six arrondissements et cent dix-neuf quartiers. La commune d'Owendo comprend deux arrondissements et treize quartiers. La commune d'Akanda comprend deux arrondissements et quatre quartiers.

Par rapport à l'organisation du système de santé du Gabon, la région sanitaire LBV-OWE-AKA comprend quatre départements sanitaires composés chacun d'un regroupement d'arrondissements :

— Département sanitaire Libreville 1 (1er et 2ème Arrondissements de Libreville plus les deux arrondissements d'Akanda) ;
— Département sanitaire Libreville 2 (3ème et 6ème Arrondissements de Libreville) ;
— Département sanitaire Libreville 3 (4ème et 5ème Arrondissements de Libreville) ;
— Département sanitaire Owendo (1er et 2ème Arrondissements d'Owendo).

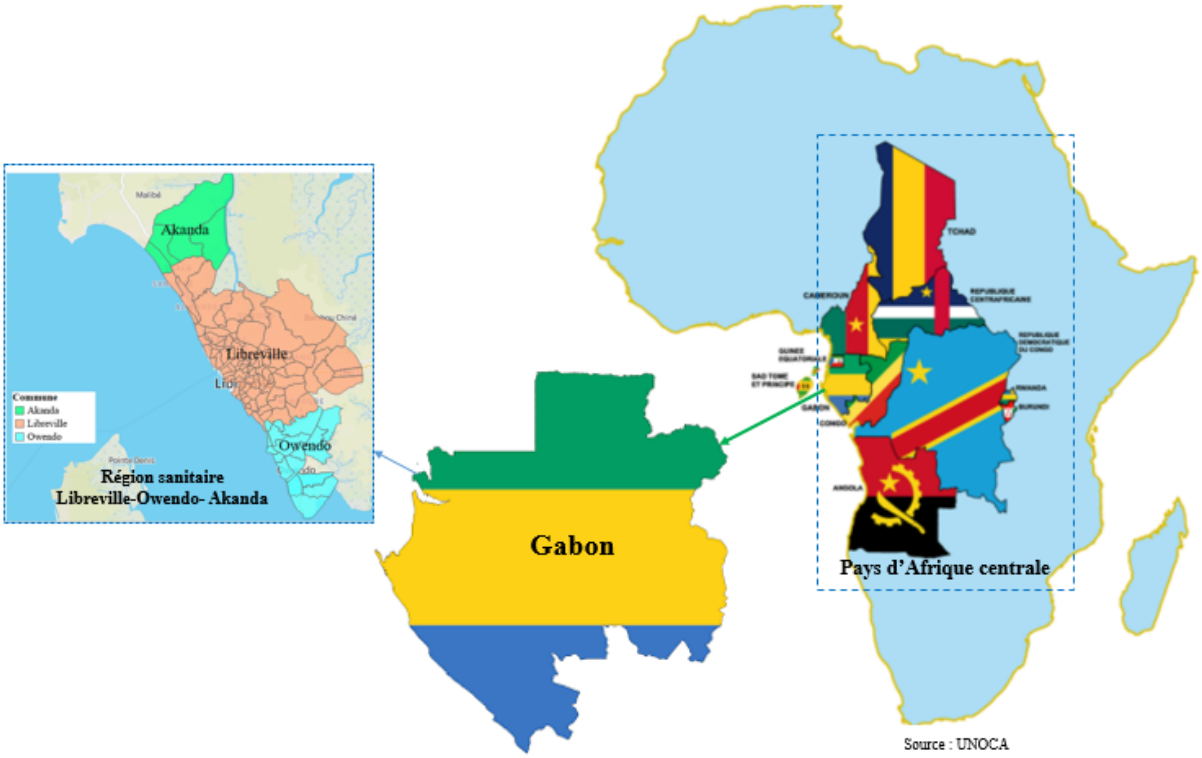


FIGURE 1 – Cadre de l'étude

LBV-OWE-AKA compte quatre centres de diagnostic et de traitement (CDT) de la tuberculose : Hôpital des Instructions des Armées Omar BONGO ONDIMBA (HIAOBO), Centre Hospitalier Universitaire de Libreville (CHUL), Prison Centrale et Hôpital Spécialisé Nkembo (HSN). Le quatrième CDT, l'hôpital spécialisé de Nkembo est l'hôpital de référence nationale spécialisé dans le traitement des malades souffrant de la tuberculose. Les données que nous utilisons dans ce travail ont été collectées dans les dossiers médicaux des personnes souffrant de tuberculose pris en charge dans ce CDT. Nous avons collecté les données dans tous les dossiers médicaux disponibles de la période 2016-2018, soit 100% de dossiers et un total de 7 968 tuberculeux recensés.

1.5 Généralités sur la tuberculose

Dans cette section, nous décrivons la cause et le mode de transmission de la tuberculose (TB), ces formes cliniques, l'algorithme de prise en charge et les différents résultats thérapeutiques. Enfin, nous présentons l'organisation du système de surveillance de la TB au Gabon.

1.5.1 Cause et mode de transmission

La TB est une maladie infectieuse à transmission interhumaine liée à une bactérie, le Mycobacterium tuberculosis ou Bacille de Koch (BK) (Jabri et al., 2016; Dombret, 2004). Cette transmission se fait essentiellement par voie aérienne, notamment lorsque qu'une personne atteinte de TB maladie tousse, parle ou éternue. La contamination se fait lors de l'inhalation des gouttelettes infectieuses (Varaine and Rich, 2014).

Selon l'histoire naturelle de la TB, son développement se fait en plusieurs phases (Denis and Perronne, 2004), la première phase de l'infection (Figure 2), nommée primo-infection tuberculeuse (PIT), fait suite au dépôt alvéolaire de bacilles tuberculeux qui, après multiplication, constituent le chancre d'inoculation (ou foyer primaire). Les bacilles se disséminent alors par voie ganglionnaire puis sanguine et constituent des foyers secondaires. Dans 90% des cas, l'infection est contenue et reste asymptomatique, on parle d'infection tuberculeuse latente (ITL). Dans les suites immédiates, 5% des cas de la PIT peuvent développer une tuberculose active avec apparition de signes cliniques. Aussi, sans traitement, la tuberculose maladie évolue vers la mort (50% des cas), vers une guérison spontanée (25% des cas) ou vers une chronicisation (25% des cas) (Denis and Perronne, 2004).

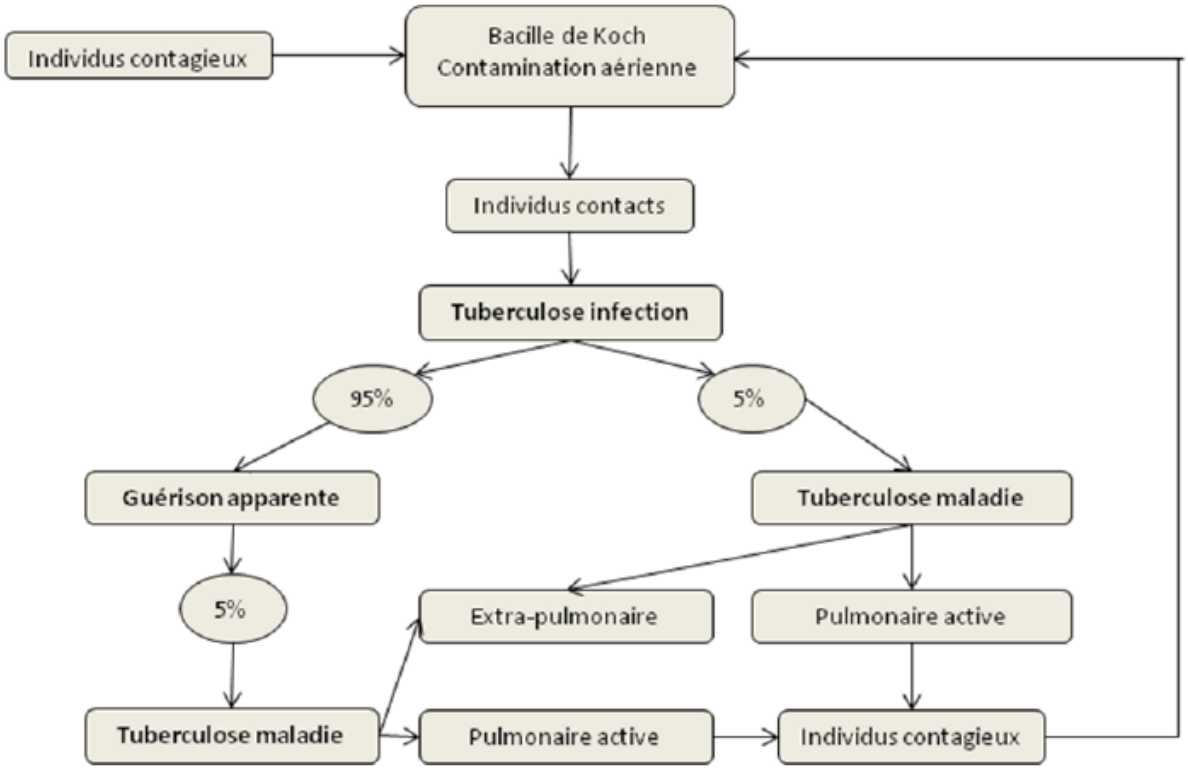


FIGURE 2 – Phases de transmission de la tuberculose

1.5.2 Différentes formes de tuberculose

Il existe plusieurs formes de tuberculose selon la localisation. Parmi celles-ci, nous pouvons citer la pulmonaire (TBP) et l'extra-pulmonaire (TBEP).

- TBP : désigne tout cas de tuberculose confirmé bactériologiquement ou diagnostiqué cliniquement dans lequel le parenchyme pulmonaire ou l'arbre trachéobronchique est touché (WHO, 2014).
- TBEP : désigne tout cas de tuberculose confirmé bactériologiquement ou diagnostiqué cliniquement dans lequel d'autres organes que les poumons sont touchés (par exemple la plèvre, les ganglions lymphatiques, l'abdomen, les voies génito-urinaires, la peau, les articulations, les os, etc.) (WHO, 2014).

Il existe également des formes de TB liés à la résistance aux antituberculeux. Autrement dit, les patients mis sous traitement antituberculeux de première intention développent une résistance aux médicaments. Ces cas sont classés par catégorie en fonction des tests de sensibilité aux médicaments menés sur des isolements cliniques confirmés de Mycobacterium tuberculosis (WHO, 2014) :

- Monorésistance : résistance à un seul antituberculeux de première intention ;
- Polyrésistance : résistance à plus d'un antituberculeux de première intention autre que l'isoniazide et la rifampicine ;
- Multirésistance : résistance au moins à l'isoniazide et à la rifampicine ;
- Ultrarésistance : résistance à une fluoroquinolone et au moins à un des trois médicaments injectables de deuxième intention (amikacine, capréomycine et kanamycine), en plus de la multirésistance ;
- Résistance à la rifampicine : une résistance à la rifampicine, détectée au moyen de méthodes phénotypiques ou génotypiques, avec ou sans résistance aux autres antituberculeux. Cette notion inclut toutes les formes de résistance à la rifampicine (monorésistance, multirésistance, polyrésistance ou ultrarésistance).

Il faut relever que les formes résistantes de la tuberculose sont plus contagieuses que la pulmonaire. Par contre, les formes extrapulmonaires ne sont pas contagieuses.

1.5.3 Prise en charge de la tuberculose

La prise en charge de la tuberculose se fait selon l'algorithme ci-après (Figure 3) (VIDAL, 2020) :

1. **Isolement** : chez un patient suspect de tuberculose pulmonaire, laryngée ou bronchique, l'isolement s'impose, ainsi que des règles particulières de déplacement et de visite. L'hospitalisation pour isolement doit être poursuivie jusqu'à la constatation d'absence de BAAR [4] sur les frottis, en moyenne entre 2 à 3 semaines de traitement

4. Bacilles Acido-Alcoolo-Résistants

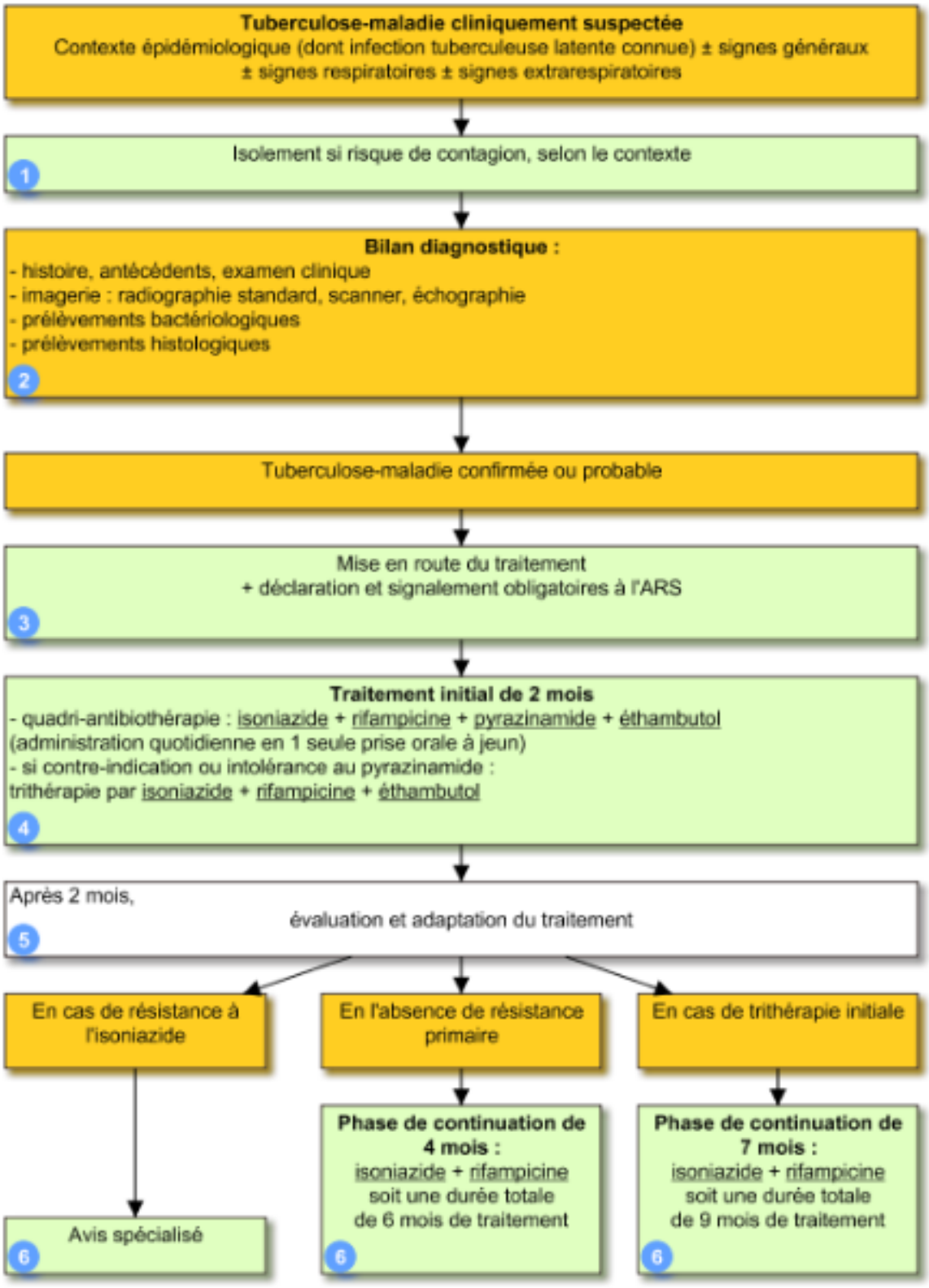


FIGURE 3 – Tuberculose pulmonaire : prise en charge (VIDAL, 2020)

en cas de BAAR sur les frottis initiaux.

2. **Bilan diagnostique** :
   — Il précise la (ou les) localisation(s) de la tuberculose. Le retentissement général,

les éléments de gravité.

— La recherche de BK [5] se fait dans les expectorations (3 jours de suite avec ou sans tubage gastrique) et, selon le cas, dans le liquide pleural, le LCR [6], les urines, la ponction d'une adénopathie. L'anatomopathologie recherche un granulome épithélioïde avec nécrose caséeuse sur les biopsies de tissus suspects d'atteinte tuberculeuse.

3. **Mise en route du traitement** :
   — Il est en général débuté sans attendre les résultats des cultures et de l'antibiogramme, mais après recherche de BAAR sur l'examen direct des frottis. Seules les méningites et certaines miliaires justifient un traitement immédiat.
   — La déclaration et le signalement de cas de tuberculose auprès de la base d'épidémiologie de lutte contre les endémies de la Direction Régionale de Santé (DRS) sont obligatoires.
4. **Médicaments de la phase initiale** : Le traitement associe quatre antibiotiques pendant deux mois : l'isoniazide à la posologie de 3 à 5 mg/kg par jour ; la rifampicine à 10 mg/kg par jour ; le pyrazinamide à 30 mg/kg par jour ; l'éthambutol à 20 mg/kg par jour. La bonne observance du traitement est essentielle pour éviter la sélection d'un mutant résistant. Des formes combinant plusieurs antibiotiques (isoniazide + rifampicine + pyrazinamide, ou isoniazide + rifampicine) améliorent l'observance. L'intolérance ou une contre-indication au pyrazinamide impose une trithérapie par isoniazide + rifampicine + éthambutol aux mêmes posologies.
5. **Évaluation et adaptation du traitement à 2 mois** : Elle dépend des résultats des cultures et d'éventuelles résistances. En cas de multirésistance du BK, le traitement sera réadapté. Il faut s'employer à obtenir l'adhésion des malades au bon suivi du traitement, car un nombre toujours trop élevé d'entre eux sont perdus de vue.
6. **Phase de continuation** :
   — En l'absence de résistance, isoniazide + rifampicine pendant 4 mois si quadrithérapie initiale, et 7 mois si trithérapie initiale.
   — En cas de résistance, se référer à un centre spécialisé.

1.5.4 Résultats thérapeutiques

Un patient mis sous traitement antituberculeux peut avoir un des résultats thérapeutiques ou issus du traitement parmi ceux présentés dans le tableau 1.1 (WHO, 2014).

5. Bacille de Koch
6. Liquide CéphaloRachidien

TABLE 1.1 – Résultats thérapeutiques pour les patients tuberculeux

Résultat thérapeutique	Définition
Guérison	Un patient atteint de tuberculose pulmonaire chez qui l'affection a été confirmée bactériologiquement en début de traitement présente des résultats négatifs (selon l'examen des frottis ou la mise en culture) au cours du dernier mois de traitement et au moins une fois auparavant.
Traitement terminé	Le patient tuberculeux a terminé le traitement sans signe d'échec. Mais on ne dispose pas de données indiquant que les résultats de l'examen des frottis ou de la mise en culture ont été négatifs au cours du dernier mois de traitement et au moins une fois auparavant, ou parce que les tests n'ont pas été réalisés, soit parce que les résultats ne sont pas disponibles.
Échec thérapeutique	Le patient tuberculeux continue de présenter des résultats positifs (selon l'examen des frottis ou la mise en culture) après cinq mois de traitement ou plus.
Décès	Le patient tuberculeux meurt pour une raison quelconque au cours du traitement ou avant de l'avoir commencé.
Perdu de vue	Le patient tuberculeux n'a pas entamé de traitement ou celui-ci a été interrompu pendant deux mois consécutifs ou plus.
Non évalué	Patient tuberculeux à qui aucun résultat thérapeutique n'a été attribué. Cette catégorie inclut les cas transférés à une autre unité de traitement (transferts sortants) et ceux dont les résultats sont inconnus de l'unité chargée de la notification.
Succès thérapeutique	Somme des patients guéris et des patients ayant terminé leur traitement.

1.5.5 Organisation du système de surveillance de la TB

Au Gabon, l'organisation du système de surveillance de la TB suit celui de la pyramide sanitaire. Cette organisation repose sur trois niveaux (Figure 4) : (i) le niveau opérationnel composé de CDT et de laboratoires qui produisent les données épidémiologiques qui seront utilisées pour la surveillance de l'infection tuberculeuse, (ii) le niveau intermédiaire dans lequel se trouve les bases d'épidémiologie et lutte contre les endémies (BELE) qui assurent la compilation des données TB transmises par les CDT et/ou laboratoire des régions sanitaires et (iii) le niveau stratégique représenté par le programme national de lutte contre la tuberculose et qui assure la compilation des données nationales. Il faut relever qu'il y a une rétro information entre les différents niveaux de cette organisation.

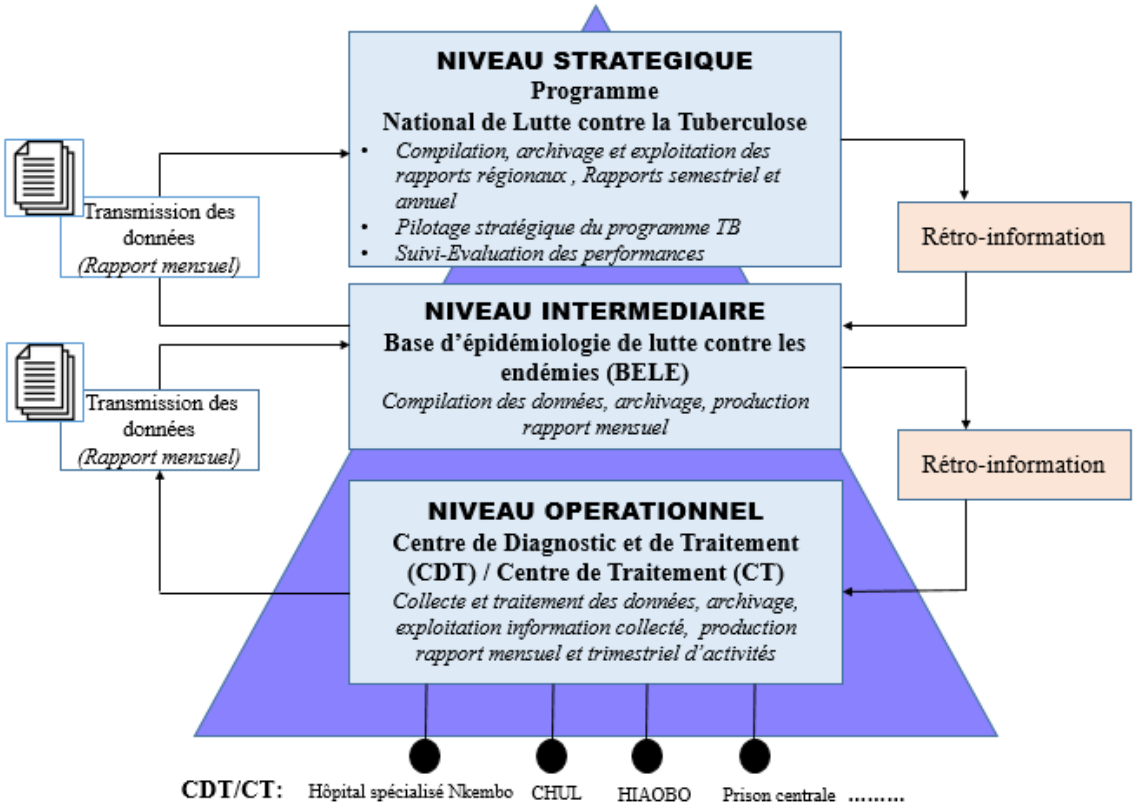


FIGURE 4 – Organisation du système de surveillance de la tuberculose

1.5.6 Gestion des données TB

Au Gabon, dans les CDT, les données des patients TB sont enregistrées dans des registres (traitement de la TB et laboratoire) et les dossiers médicaux des malades, tous au format papier. Le système de rapportage est trimestriel pour les CDT et mensuel pour les Laboratoires. Une fois le rapport d'activités produit par le responsable du CDT ou du laboratoire, la version papier est ensuite transmise au chef de base d'épidémiologie et de lutte contre les endémies (BELE) qui compile les données et produit un rapport qui est par la suite envoyé au PNLT. Le bureau Suivi/Évaluation du PNLT effectue la compilation manuelle de tous les rapports et produit les rapports semestriel et annuel de la tuberculose au Gabon.

1.6 Plan de la thèse

Le mémoire de thèse est organisé en deux grands chapitres (Figure 5). Le premier chapitre est consacré à l'état de l'art et s'articule autour de quatre sections.

Section 2.1 : Qualité des données en santé publique

Cette section présente les concepts généraux sur la qualité des données, suivie d'un état de l'art sur la qualité des données dans les systèmes de surveillance épidémiologique

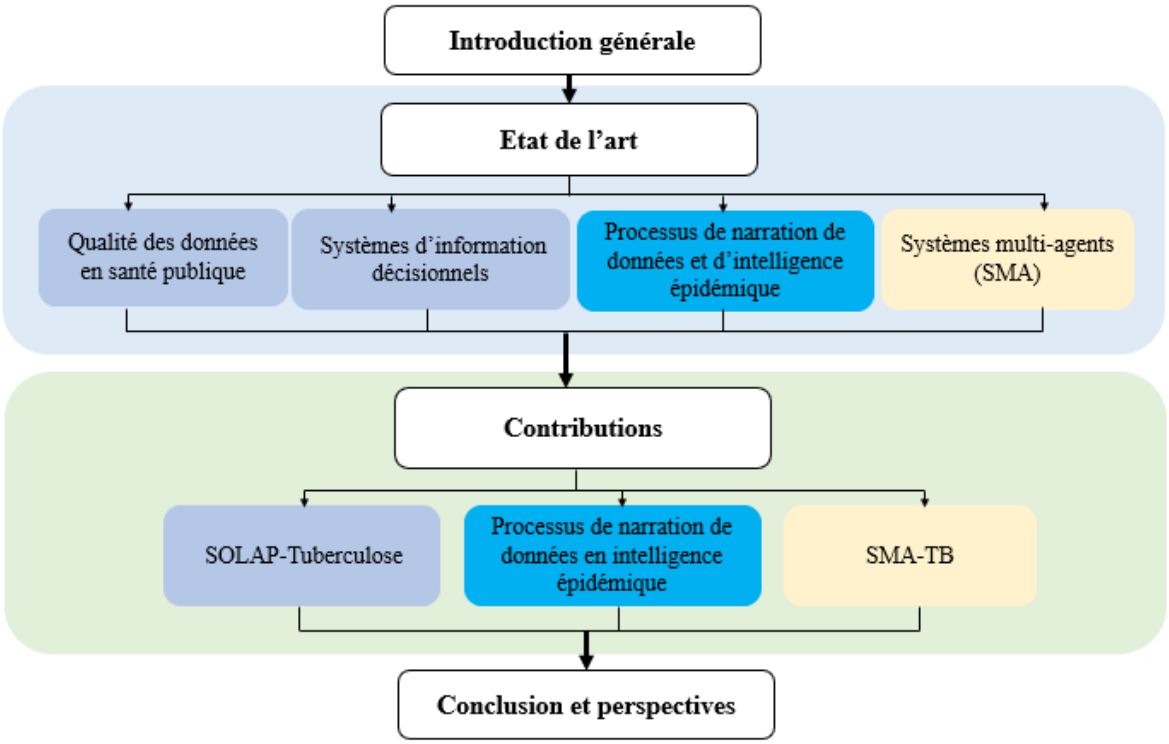


FIGURE 5 – Organisation du mémoire de thèse

des maladies.

Section 2.2 : Systèmes d'information décisionnels

Cette section constitue une présentation générale des concepts de base sur les systèmes d'information décisionnels OLAP et Spatial-OLAP ou SOLAP. Elle présente également une étude sur les SOLAP appliqués à la surveillance épidémiologique des maladies.

Section 2.3 : Processus de narration de données et d'intelligence épidémique

Cette section constitue une revue de la littérature sur les processus de narration de données et d'intelligence épidémique. Une étude de la narration des données en santé publique est également réalisée.

Section 2.4 : Système multi-agents

Cette section constitue une présentation générale des concepts de base sur les systèmes multi-agents (SMA). Elle présente également une étude sur l'utilisation des systèmes multi-agents pour la modélisation et la simulation de la transmission inter-individus des maladies infectieuses, en faisant un focus sur la tuberculose.

Le deuxième chapitre est consacré à la présentation des différentes contributions proposées dans cette thèse. Cette partie se compose de trois sections.

Section 3.1 : Système d'information géographique décisionnel pour la surveillance de la tuberculose

Cette section présente de façon détaillée le processus de conception du système d'information géographique décisionnel pour la surveillance épidémiologique de la tuberculose

(SOLAP-TB). À chaque étape de ce processus, des résultats sont présentés.

Section 3.2 : Processus de narration de données en intelligence épidémique Cette section présente une description détaillée du processus de narration de données en intelligence épidémique (P-N-D-I-E) proposé. Une application de production d'une narration de données sur la pandémie de la tuberculose au Gabon qui s'appuie sur ce P-N-D-I-E et le SOLAP-TB proposé est également présentée.

Section 3.3 : Système multi-agents de la tuberculose (SMA-TB)

Cette section présente une description détaillée du modèle de système multi-agents de la tuberculose (SMA-TB) proposé. Aussi, les résultats des simulations de différentes politiques sanitaires réalisées avec le simulateur SMA-TB conçu y sont également présentés et discutés.

Chapitre 2

État de l'art

Ce chapitre présente les concepts généraux sur la qualité des données (section 2.1), les systèmes d'information décisionnels (section 2.2), les processus de narration de données et d'intelligence épidémique (section 2.3), ainsi que sur les systèmes multi-agents (section 2.4). Aussi, dans chacune de ces sections, en s'appuyant sur la littérature, nous étudions respectivement (i) la qualité des données dans les systèmes de surveillance des maladies, (ii) les systèmes d'information décisionnels pour la surveillance de la tuberculose, (iii) la narration de données en santé publique et (iv) les systèmes multi-agents de simulation de la propagation inter-individus de la tuberculose.

2.1 Qualité des données en santé publique

Les problèmes de qualité des données (tels que les erreurs typographiques, les doublons, les incohérences, les valeurs manquantes, incomplètes, incertaines, obsolètes, ou peu fiables) se posent de façon récurrente dans tous les systèmes d'information, bases et entrepôts de données et pour tous les domaines d'application (Akoka et al., 2008). Dans le domaine de la santé publique, en ce qui concerne les systèmes de surveillance des maladies (COVID-19, paludisme, tuberculose, etc.), il est important que les données utilisées pour le calcul des indicateurs soient d'abord évaluées en termes de qualité. Parce que, le manque de fiabilité des données sanitaires rend la planification et la prise de décisions extrêmement difficiles. En effet, si l'analyse des données et la prise de décision peuvent être réalisées sur des données inexactes, incomplètes, ambiguës et de qualité médiocre, on peut légitimement s'interroger sur le sens à donner aux résultats de ces analyses et, à juste titre, remettre en cause la qualité des connaissances découvertes à partir de ces données ainsi que le bien-fondé des décisions prises (Berti-Equille, 2007).

Dans cette section, nous réalisons un état de l'art sur la qualité des données en santé publique. Elle est organisée comme suit : à la sous-section 2.1.1, nous passons d'abord en revue les concepts clés sur la qualité des données. Ensuite, à la section 2.1.2, nous présentons les travaux qui ont porté sur la qualité des données dans les systèmes de surveillance des maladies. Nous terminons par une conclusion.

2.1.1 Concepts de la qualité des données

2.1.1.1 Notion de qualité des données

La qualité des données est un des principaux domaines de recherche actuel, en particulier dans le cadre des entrepôts de données (Fruytier, 2009). Elle est généralement définie comme l'adéquation des données à un cas d'utilisation, en soulignant sa nature relative et dynamique, dont le contexte est déterminé par l'utilisation des données et les exigences qui en dépendent, et qui peuvent changer au fil du temps, c'est-à-dire en fonction de l'accumulation progressive des données dans les bases de données et de l'évolution des exigences de qualité des données (Nikiforova, 2020). Les problèmes de qualité des données sont mesurés à l'aide des métriques qui permettent d'évaluer les dimensions de la qualité des données.

À la sous-section 2.1.1.2, nous présentons les dimensions de la qualité des données.

2.1.1.2 Dimensions de la qualité des données

De nombreux travaux, par exemple, (Wand and Wang, 1996; Scannapieco et al., 2005; Berti-Équille, 2006; Akoka et al., 2008) ont montré que la qualité des données est un concept multidimensionnel. Car, elle est caractérisée par un ensemble de dimensions qui décrivent les données fournies aux utilisateurs. Une dimension de la qualité des données correspond à une caractéristique d'une donnée qui permet de la classifier et d'en définir les besoins au niveau de sa qualité (Goulet, 2012). Les dimensions qualité sont mesurées par des métriques qui sont utilisées pour déterminer l'utilité et la pertinence des données, et aident à séparer les données de haute qualité des données de très faible qualité (Berti-Équille, 2004). Selon (IEEE, 1998), une métrique est une fonction qui associe une dimension de qualité à une valeur numérique, ce qui permet d'interpréter la réalisation d'une dimension.

Dans la littérature, de nombreuses dimensions qualité ont été proposées pour caractériser les multiples facettes de la qualité des données utilisées par les entreprises et les organisations pour la prise de décisions stratégiques. Parmi ces dimensions, nous pouvons citer : la complétude, l'unicité, la validité, la fraîcheur et la cohérence. Aux paragraphes ci-après, nous présentons chacune de ces dimensions.

2.1.1.2.1 Complétude : La complétude mesure principalement des taux de déficit (données manquantes par rapport au terrain nominal de référence) ou d'excès (données présentes dans le jeu de données, mais manquantes ou indéterminées sur le terrain nominal) (Akoka et al., 2008). La complétude se mesure en détectant le taux de valeurs manquantes dans la base de données (Berti-Équille, 2004). Par exemple, le pourcentage de données manquantes pour chacune des trois variables obligatoires lors de l'enregistre-

ment prénatal (âge, parité [1] et hémoglobine de base) (Amoakoh-Coleman et al., 2015).

2.1.1.2.2 Unicité : Cette dimension qualité définit à quel point les données évitent les redondances (des données sont redondantes si elles décrivent un même objet du monde réel) (Barrau et al., 2016). Elle a pour objectif de s'assurer que les données ne sont pas dupliquées dans une base de données. L'unicité se mesure en détectant le nombre d'enregistrements avec une entrée en double dans une table (Ehrlinger and Wöß, 2022). Par exemple, lors de la prise en charge d'un patient tuberculeux, le personnel soignant ne vérifie pas toujours si le patient existe déjà dans la base de données (c'est dire qu'il a déjà été pris en charge dans le centre de diagnostic et de traitement) et peut-être amené à recréer à tort ce patient comme « nouveau patient ». Il en résulte des doublons de patients tuberculeux. Une manière de pallier ce problème pourrait être de créer un identifiant unique (par exemple, le numéro d'assurance maladie de garantie sociale) pour les patients.

2.1.1.2.3 Validité : La validité des données est définie comme le pourcentage de données dont les valeurs se situent dans leur domaine respectif de valeurs admissibles (Berti-Equille, 2007). Par exemple, étant donné un champ « âge patient » dans une table « patient », si ce champ contient la valeur «3ème âge », il n'est donc pas valide.

La validité se mesure en détectant le nombre de valeurs d'attributs qui ne sont pas conformes au type de données d'une colonne (Ehrlinger and Wöß, 2022).

2.1.1.2.4 Fraîcheur : Elle désigne l'ensemble des facteurs qui capturent les caractères récents et d'actualité d'une donnée entre l'instant où elle a été extraite ou créée dans la base de données et l'instant où elle est présentée à l'utilisateur (Berti-Équille, 2006). Par exemple, une étude porte sur le profil épidémiologique d'une maladie (par exemple paludisme, VIH, tuberculose) en 2020 dans un pays. Mais, les données disponibles dans la base de données nationale pour cette maladie datent de 1996. Ces données sont obsolètes pour cette étude.

La fraîcheur des données se mesure par une comparaison entre la date de saisie et la date courante (Berti-Équille, 2004).

2.1.1.2.5 Cohérence : Cette dimension assure l'absence d'informations conflictuelles au sein d'un même objet ou entre objets différents. Par exemple, la date de naissance d'un enfant antérieure à la date de naissance des parents (Berti-Equille, 2012). La cohérence est liée à un ensemble de contraintes (ou règles métier). Elle se mesure par rapport à un ensemble de contraintes en détectant les données de la base de données qui ne les satisfont pas (Berti-Équille, 2004).

1. Nombre d'accouchements précédents

Pour sélectionner les dimensions qualité à évaluer, les critères permettant de procéder à cette sélection doivent prendre en considération certains facteurs, tels que le cycle de vie des données, le type de données manipulées et le type d'entreprise dans lequel nous œuvrons (Goulet, 2012). Dans cette étude, nous avons retenu d'évaluer la qualité des données recueillies dans le système de surveillance épidémiologique de l'infection tuberculeuse selon trois dimensions qualité : la complétude, l'unicité et la validité des données. Plusieurs raisons ont guidé ce choix :

D'abord, concernant la complétude des données, cette dimension est l'une des dimensions qualité les plus fréquemment évaluée pour la qualité des données dans la littérature existante sur les soins de santé (Weiskopf and Weng, 2013). Car, les problèmes liés à la complétude des données peuvent avoir de graves conséquences dans le domaine des soins de santé (Liu et al., 2017). À titre d'exemple, l'incomplétude des données entraîne une incertitude importante dans les indicateurs de santé tels que les taux d'incidence, de prévalence et de létalité d'une maladie (par exemple : la COVID-19, le VIH, le paludisme, la tuberculose).

Pour ce qui est de l'unicité des données, l'un des gros problèmes dans la gestion d'une base de données (par exemple patients suivi dans un hôpital, clients d'une entreprise) est la présence de doublons. Il est donc est important d'identifier ces doublons dans la base de données. Car, par exemple, dans le domaine de la santé, la non-élimination de ces patients en doublon infectés par une maladie dans la base de données de surveillance épidémiologique entraîne une fausse augmentation du nombre de cas enregistrés.

Enfin, concernant la validité des données, nous avons retenu d'évaluer la validité des divers types de données affectées aux différentes variables (par exemple le genre, le sexe, l'âge, le type de tuberculose, le résultat thérapeutique) qui caractérisent le patient tuberculeux.

Lorsque différents problèmes de qualité sont détectés dans une base de données, il faut procéder à l'amélioration de ces problèmes de qualité afin de permettre aux responsables d'une organisation ou entreprise de prendre des décisions avec des données de bonne qualité. La sous-section 2.1.1.3 présente quelques techniques d'amélioration de la qualité basées sur les données.

2.1.1.3 Techniques d'amélioration de la qualité des données

Les techniques d'amélioration de la qualité des données peuvent être soient basées sur les processus ou sur les données. Pour celle qui est basée sur les données qui nous intéressent dans cette étude, la liste des techniques (non exhaustive) est la suivante (Batini et al., 2009) :

— L'acquisition de nouvelles données, qui améliore les données en acquérant des données de meilleure qualité pour remplacer les valeurs qui posent des problèmes de qualité ;

— La standardisation (ou normalisation), qui remplace ou complète les valeurs de données non standard par des valeurs correspondantes conformes à la norme ;
— Le couplage d'enregistrements, qui permet d'identifier que les représentations de données dans deux (ou plusieurs) tables qui pourraient se référer au même objet du monde réel ;
— L'intégration des données et des schémas, qui définit une vue unifiée des données fournies par des sources de données hétérogènes. L'intégration a pour principal objectif de permettre à un utilisateur d'accéder aux données stockées par des sources de données hétérogènes grâce à une vue unifiée de ces données ;
— La sélection des sources, qui permet de sélectionner les sources de données en fonction de la qualité de leurs données ;
— La localisation et la correction des erreurs, qui identifient et éliminent les erreurs de qualité des données en détectant les enregistrements qui ne satisfont pas à un ensemble donné de règles de qualité. Ces techniques sont principalement utilisées dans le domaine de la statistique ;

Pour détecter et corriger la qualité des données stockées dans un entrepôt de données, différentes approches peuvent être utilisées. Ces différentes approches sont présentées à la sous-section 2.1.1.4.

2.1.1.4 Approches pour détecter et corriger les problèmes de qualité des données

Les approches de détection et correction des problèmes de qualité des données stockées dans un entrepôt ou une base de données peuvent être classées en quatre grands groupes (figure 6) (Berti-Équille, 2006) :

— Les approches préventives centrées sur l'ingénierie des systèmes d'information et le contrôle des processus avec des techniques permettant d'évaluer la qualité des modèles conceptuels, la qualité des développements logiciels et celle des processus employés pour le traitement des données ;
— Les approches diagnostiques centrées sur des méthodes statistiques, d'analyse et de fouille de données exploratoire permettant de détecter des anomalies sur les données ;
— Les approches correctrices centrées sur des techniques de nettoyage et consolidation des données et utilisant des langages de manipulation des données étendus et des outils d'extraction et de transformation de données (ETL- Extraction-Transformation-Loading) ;
— Les approches adaptatives ou actives appliquées généralement lors de la médiation ou de l'intégration des données : elles sont centrées sur l'adaptation des traitements (requêtes ou opération de nettoyage sur les données de telle façon que ceux-ci incluent à l'exécution en temps réel la vérification de contraintes sur la qualité des

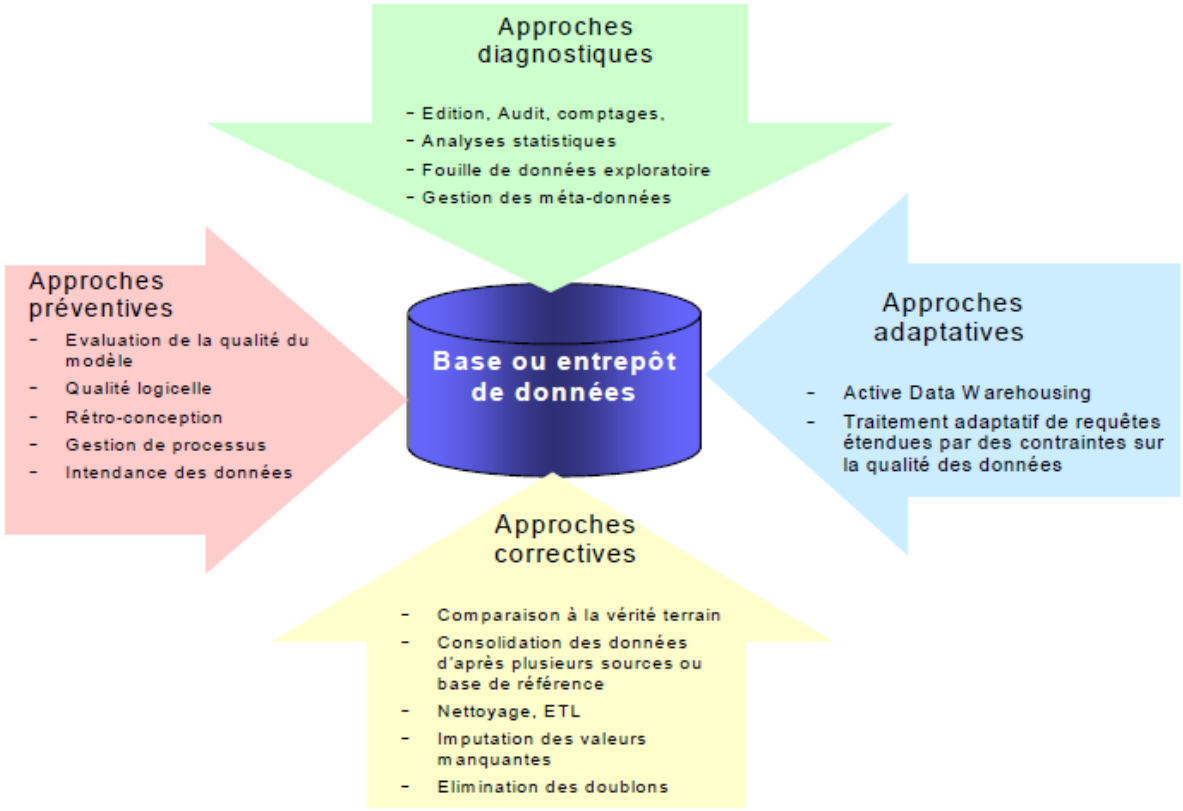


FIGURE 6 – Approches d'évaluation et du contrôle de la qualité des données (Berti-Équille, 2006)

données.

Dans ce travail, nous avons utilisé une approche correctrice en utilisant un outil ETL. La sous-section 2.1.1.5 présente les processus ETL et quelques outils ETL.

2.1.1.5 Processus Extraction-Transformation-Chargement (ETL) et outils ETL

Le processus Extraction-Transformation-Chargement (ETL) est une partie cruciale du processus d'entreposage de données où la plupart des nettoyages et des traitements de données sont effectués (Souibgui et al., 2019). Selon (Vassiliadis and Simitsis, 2009), la description d'un processus ETL se résume en trois étapes (Figure 7) [2] :

— L'extraction (Extract) : les données sont extraites des sources qui peuvent être structurées (relationnelles) ou non structurées (des pages web, des fichiers tabulaires, des flux de données, etc.).

— La transformation (Transform) : Les données sont propagées dans un espace de stockage temporaire appelé "Data Staging Area" dans lequel des opérations de transformation, d'homogénéisation, de nettoyage et de filtrage sont mises en place.

— Le chargement (Load) : Les données sont chargées dans l'entrepôt de données. On parle de matérialisation des données.

2. (https ://www.keboola.com

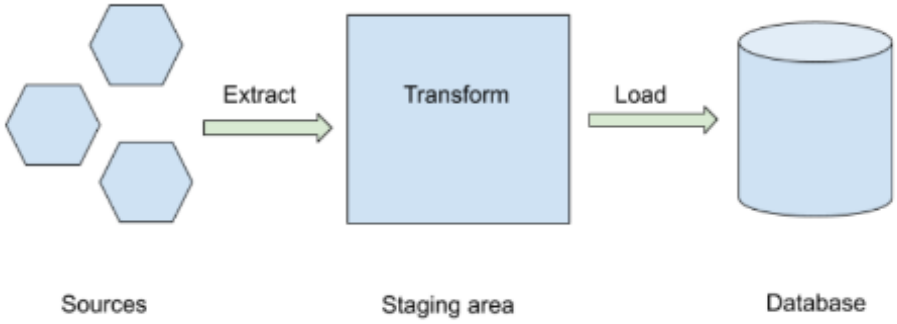


FIGURE 7 – Processus ETL

Dans un système décisionnel, les étapes du processus ETL sont souvent automatisés en utilisant un outil ETL. Ce type d'outil permet de collecter, transformer et consolider les données de manière automatisée. Les avantages de l'utilisation d'un tel outil dans un système décisionnel sont de (Crozat, 2016) :

- Structurer et de rassembler l'ensemble des morceaux de code nécessaire aux transferts et aux transformations des données ;
- Intégrer la gestion des métadonnées ;
- Intégrer la gestion des erreurs ;
- Disposer d'API (interface de programmation d'application) dédiées (connexion, import/export...) d'accès aux données (CSV, Base de Données, XML...).

Il existe actuellement de nombreux outils ETL. Parmi ces outils, certains sont propriétaires (IBM Information Server, SAS Data Integration Studio, Oracle Warehouse Builder, Sap Business Objects Data Integration, etc.) et d'autres sont Open Source (Talend Open Studio, Pentaho Data Integration, Clover ETL, Alterix, etc.). En outre, la plupart des outils de l'informatique décisionnelle (BI) comme Tableau, Power BI, etc. embarquent désormais leur propre ETL. Par exemple, Power BI embarque l'ETL "Dataflow" et Tableau l'ETL "Tableau Prep". Concernant Tableau Prep de l'outil BI Tableau que nous utilisons dans notre contribution SOLAP de la tuberculose, il est vraiment très basique au niveau des transformations. Car, il s'agit d'un outil de préparation de données dont l'unique vocation est de servir les visualisations sous Tableau. C'est pourquoi, nous avons opté d'utiliser Talend Open Studio qui a des fonctionnalités plus avancées pour le traitement de données. Ce choix a priori a été guidé par le fait que nous avons utilisé cet ETL dans le cadre académique.

Talend Open Studio (TOS) est un outil ETL open source convivial basé sur Eclipse. Il utilise le langage Perl ou Javascript pour définir ses propres composants. Il est hautement évolutif et fonctionne sous Windows, Unix et Linux. Il est spécialement utilisé pour l'intégration de données, la qualité des données, la gestion des données (Kherdekar and Metkewar, 2016). Il s'inspire d'outils ETL propriétaires tels que DataStage ou Informatica et comprend un composant de mapping graphique Talend tMap (Pushkarev et al., 2010).

La figure 8 illustre l'architecture fonctionnelle d'intégration des données de Talend Open Studio [3]. Cette architecture a été décrite dans (Kumar et al., 2019) :

- Le bloc Clients comprend un ou plusieurs Talend Studio(s) et des navigateurs Web qui peuvent se trouver sur la même machine ou sur des machines différentes. Depuis Talend Open Studio, les processus d'intégration de données quel que soit le niveau de complexité des données et des processus peuvent être réalisés. Talend Studio permet de travailler sur n'importe quel projet pour les processus autorisés. Aussi, depuis le navigateur, il est possible de se connecter au centre d'administration Talend à distance via un protocole HTTP sécurisé ;
- Le bloc Server comprend le serveur d'application web et le Talend Administration Center, qui permet la gestion et l'administration de tous les projets. Les méta-données d'administration (comptes utilisateurs, droits d'accès et autorisations de projets par exemple) sont stockées dans la base de données d'administration. Les données des éléments du projet (Jobs, Business Models et Routines par exemple) sont stockées dans le serveur SVN ou Git ;
- Le bloc Référentiels comprend le serveur SVN ou Git et le référentiel de données. Le serveur SVN ou Git est utilisé pour organiser tous les éléments de projets tels que les Jobs et les Business Models partagés entre différents utilisateurs finaux. Il est accessible depuis Talend Open Studio pour développer des éléments de projets, et depuis Talend Administration Center pour publier, déployer et surveiller les éléments de projets ;
- Les blocs Talend Execution Server comprennent un ou plusieurs serveurs d'exécution, qui seront déployés au sein du système d'information.
- Talend Jobs est déployé sur les serveurs de jobs via le Job Conductor du Talend Administration Center pour être exécuté à une heure, une date ou un événement programmé.
- Le bloc Bases de données comprend les bases de données Administration, Audit et Monitoring. La base de données Talend Administration est utilisée pour gérer les comptes utilisateurs, les droits d'accès, les autorisations de projets, etc. La base de données Audit est utilisée pour évaluer les différents aspects des Jobs mis en œuvre dans les projets développés dans Talend Open Studio dans le but de fournir des facteurs quantitatifs et qualitatifs solides pour l'aide à la décision orientée processus.

Après le passage en revue des concepts les plus importants sur la qualité des données, des méthodes et des outils de traitement de la qualité des données, à la section 2.1.2, nous étudions la qualité des données dans les systèmes de surveillance des maladies. L'objectif est de ressortir les différents problèmes de qualité des données retrouvés dans ces systèmes et les méthodes utilisées pour les traiter.

3. https ://help.talend.com/

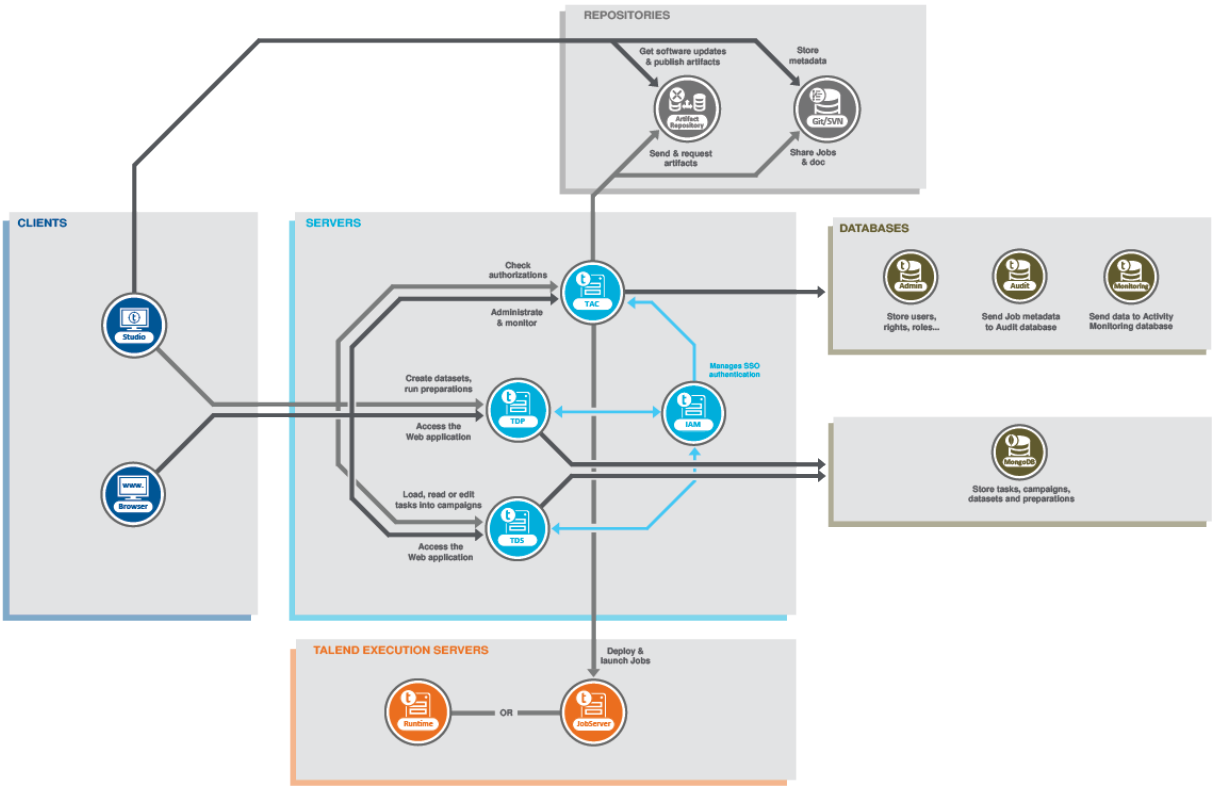


FIGURE 8 – Architecture fonctionnelle d'intégration de données Talend

2.1.2 Étude de la qualité des données dans les systèmes de surveillance des maladies

Avant d'utiliser les données pour la prise de décision stratégique dans la lutte contre les maladies, il faut d'abord s'assurer que ces données sont de très bonne qualité. Dans cette section, nous présentons des études qui ont été menées sur la qualité des données dans les systèmes d'information des programmes de santé (tuberculose, vaccination, cancer, VIH).

Dès 1966, un rapport technique de l'OMS (OMS, 1966) relevait déjà un problème de la qualité des statistiques sanitaires dans les pays en développement. Cette mauvaise qualité des statistiques sanitaires était liée au manque de systèmes normalisés pour le classement des données, la présentation des statistiques et le calcul des taux et indices. Selon le même rapport, les principaux problèmes à résoudre pour obtenir des statistiques sanitaires de bonne qualité se posaient au niveau de la collecte des données de base.

Plus tard, des études d'évaluation de la qualité des données dans les systèmes d'information des programmes de santé (tuberculose, VIH, cancer, vaccination) ont été réalisées.

Dans (Majerovich et al., 2017), l'étude a porté sur l'évaluation de la qualité des données de surveillance de l'infection tuberculeuse latente recueillies par l'intermédiaire du Système d'Information sur la Santé Publique intégré (SISP-i) dans la région de Peel, en Ontario. Les dimensions de la qualité des données évaluées étaient la complétude et la validité des données. Les résultats ont révélé que sur 6 576 registres des cas de tuberculose du SISP-i

évalués, plus de la moitié des champs dédiés aux facteurs de risque de la tuberculose étaient vides ou portaient la mention « inconnu ». Aussi, une comparaison de 192 dossiers médicaux aux registres correspondants du SISP-i a permis d'identifier des erreurs de codage dans plus de 40% des champs dédiés aux facteurs de risque de la tuberculose et l'achèvement du traitement décrit dans le SISP-i (20%) était inférieur aux données obtenues lors d'une enquête téléphonique de suivi des cas (50%). À la suite de cette évaluation, pour améliorer la qualité des données dans le SISP-i, il a été proposé entre autre la standardisation du processus de saisie des données.

En Afrique, selon de nombreux rapports (Holmes et al., 2013), les données tirées des registres et des archives hospitalières ainsi que les données brutes collectées spécifiquement à des fins de recherche peuvent être difficiles à analyser. Car, elles ne répondent pas aux normes de qualité. Par exemple, une étude relative à la prévention systématique de la transmission du VIH de la mère à l'enfant (PTME) a révélé un problème de complétude des données. En effet, 50% des données nécessaires ne figuraient pas dans les rapports adressés au système d'information sanitaire de district dans le Kwazulu-Natal, en Afrique du Sud (Mate et al., 2009).

Dans une autre étude réalisée au Mali par (Beyeme-Ondoua, 2007), la qualité des données nationales du cancer colorectal [4] a été évaluée. La dimension de la qualité des données évaluée était la complétude des données. Les résultats ont révélé que sur 556 206 séjours hospitaliers, le diagnostic principal était absent pour 134 séjours (0,02%) de l'ensemble des séjours) et que la durée de l'hospitalisation n'était pas mentionnée dans 284 731 (51,19%) séjours.

Dans une étude menée en Côte d'Ivoire en 2015, (Vroh et al., 2015) ont évalué la qualité des données de vaccination chez les enfants de 0-11 mois. Trente (N=30) districts sanitaires et quatre-vingt-huit (N=88) centres de santé ont été enquêtés. La dimension qualité évaluée était la complétude des données. La méthodologie utilisée a consisté au recomptage du nombre d'enfants ayant reçu trois doses du vaccin DTC-HepB-Hib, à partir des sources de données existantes (rapports du Programme Élargi de Vaccination des centres de santé, outil électronique d'enregistrement des vaccinations) aux différents niveaux de la pyramide sanitaire (centres santé et district). L'étude a permis de montrer que le nombre d'enfants ayant reçu trois doses de DTC-HepB-Hib variait d'un support à l'autre. Par exemple, sur les outils primaires d'enregistrement, 39% des vaccinations recomptées sur les fiches de pointage ne figuraient pas dans le registre de vaccination. Pour améliorer la qualité des données de vaccination chez les enfants, les auteurs proposent au Programme Élargie de Vaccination (PEV) de faire des supervisions afin de s'assurer que les supports de collecte de données sont bien utilisés et archivés. Aussi, au niveau district, il propose que soit mis en place des mécanismes de validation des données sur la base des

4. Le cancer colorectal, ou cancer du côlon-rectum, est le 3e cancer le plus fréquent chez l'homme après ceux de la prostate et du poumon.

outils standardisés.

Enfin, selon l'étude de (Koumamba et al., 2020), les données produites par les structures de santé dans l'actuel système d'information sanitaire (SIS) du Gabon sont de très faible qualité avec une complétude estimée à 39%. Dans le cas spécifique du système de surveillance de l'infection tuberculeuse, le rapport de la revue du programme national de lutte contre la tuberculose (PNLT) 2018 du Gabon a relevé que la base de données du PNLT comportait beaucoup d'insuffisances en termes de complétude.

2.1.3 Conclusion

Dans cette section, nous avons présenté les principaux concepts sur la qualité des données, les techniques de traitement de qualité des données et les outils qui permettent de traiter ces problèmes de qualité. Ensuite, nous avons réalisé une étude sur la qualité des données dans les systèmes de surveillance des maladies. L'objectif était de ressortir les principaux problèmes de qualité qu'on y rencontre et les techniques utilisées pour les corriger. Cette étude montre que la dimension qualité la plus évaluée dans ces systèmes est la complétude des données. Ce résultat rejoint celui retrouvé dans l'étude de (Weiskopf and Weng, 2013) qui a montré que la complétude était la dimension de la qualité des données la plus souvent évaluée dans la littérature existante sur les soins de santé.

Pour améliorer cette complétude des données, certains auteurs proposent des pistes de solutions. Parmi celles-ci, il y a la standardisation des processus de saisie de données (Vroh et al., 2015; Majerovich et al., 2017) et la validation des données sur la base des outils standardisés (Vroh et al., 2015).

Dans le cadre de notre étude, tout comme (Vroh et al., 2015) et (Majerovich et al., 2017), nous proposons la standardisation du formulaire d'enregistrement des personnes infectées par la tuberculeuse. Ce formulaire intègre toutes les variables utiles pour la déclaration des cas de tuberculose maladie et sera accessible aux Centres de diagnostic et de traitement de la tuberculose via une plate-forme Web.

2.2 Systèmes d'information décisionnels

Dans cette section, nous décrivons les systèmes d'information décisionnels ou SID (OLAP et SOLAP) et réalisons une revue de la littérature sur les SOLAP appliqués à la santé publique dans le domaine de la surveillance de maladies.

2.2.1 Systèmes OLAP (Online Analytical Processing)

L'OLAP (Online Analytical Processing) est une technique d'analyse qui a été élaborée en 1993 par Edgar Frank Codd pour sélectionner les données stockées dans une base de

donnée selon des multiples critères, d'où la notion de « multidimensionnelle ». Dans cette section, nous décrivons le système OLAP.

2.2.1.1 Architecture générale d'un système OLAP

Nous présentons l'architecture générale d'un système OLAP à la figure 9. Cette architecture est décomposée en quatre niveaux (Naoum, 2006) : (i) les sources de données, (ii) le système de stockage de données (entrepôt de données), (iii) le serveur OLAP et (iv) les outils d'analyse. Chaque composant de cette architecture est décrit ci-après :

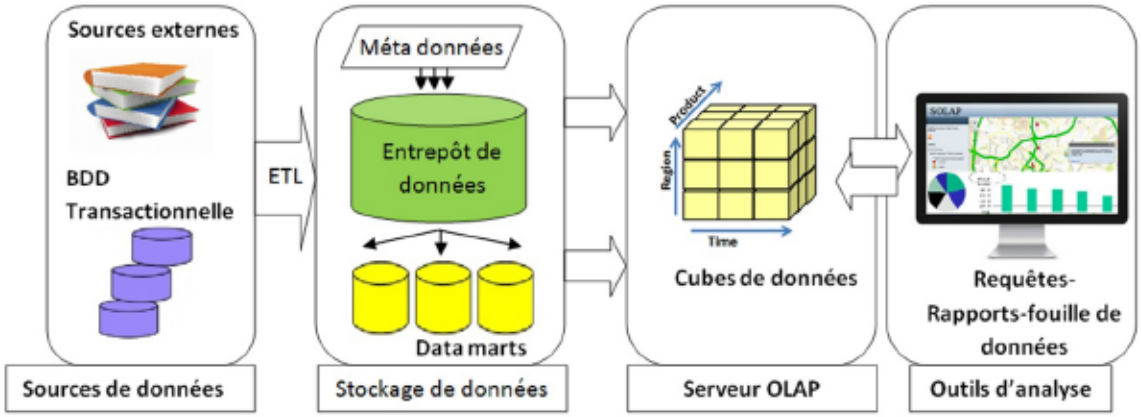


FIGURE 9 – Architecture d'un système d'information décisionnel (Charef, 2019)

Sources de données :

Les sources de données potentielles dans un système d'information décisionnel sont diverses. Il peut s'agir de bases de données, de fichiers sous la forme de tableaux, les rapports, etc. Par exemple, l'entrepôt de données d'un système d'information décisionnel appliqué à la surveillance épidémiologique d'une maladie infectieuse comme la tuberculose peut être alimenté par des données provenant de sources hétérogènes (par exemple les données cliniques et administratives des patients, les données de laboratoire, etc.) avec des formats différents (par exemple les fichiers, les base de données, etc.). Il peut également s'agir des données démographiques (par exemple la population de notre zone d'étude) et géographiques (par exemple les limites administratives d'une zone d'étude : quartiers, arrondissements, etc.). Aussi, pour que ces données hétérogènes soient exploitables pour l'aide à la décision, elles doivent suivre un processus d'extraction, de transformation et de chargement (ETL). Par exemple, dans le système décisionnel qui a été proposé par (Bouba, 2015), l'ETL Talend a été utilisé pour nettoyer (suppression des données incohérentes), homogénéiser (définition de format commun) et organiser (structure multidimensionnelle) les données issues de différentes sources. De même, (Younsi, 2016) a intégré un processus ETL dans son Système d'Information Décisionnel Spatio-temporel (SYDSEP) pour le

suivi et la surveillance du phénomène de propagation de l'épidémie de la grippe saisonnière au sein de la population de la ville d'Oran en Algérie.

Entrepôt de données :

En 1995, Inmon a défini un entrepôt de données (ou data warehouse en anglais) comme « une collection de données orientées vers un sujet, intégrées, variables dans le temps et non volatiles, destinée à soutenir le processus décisionnel » (Inmon, 1995). Donc, dans un entrepôt de données (Charef, 2019) : (i) les données sont organisées selon un sujet d'analyse en tenant compte des besoins analytiques de l'entreprise, (ii) les données hétérogènes issues de différentes sources feront l'objet d'une intégration dans un seul espace de stockage, (iii) les données ne sont ni modifiables ni supprimables (iv) du fait que les données sont non volatiles, des intervalles de temps leur sont associés. En outre, les magasins de données constituent un extrait de l'entrepôt pour répondre à un besoin d'analyse particulier (Teste, 2000).

Serveur OLAP :

Le serveur OLAP a pour rôle d'extraire les données via des requêtes SQL et d'interpréter ces données selon une vue multidimensionnelle avant de les présenter au module client. Aussi, il permet d'optimiser le temps de réponse des requêtes, pré-calcule les différents agrégats en utilisant des fonctions d'agrégation classiques (SUM, MIN, MAX, AVG ou COUNT) (Bimonte et al., 2007).

Outils d'analyse :

La restitution est assurée par les outils d'analyse. En effet, ils permettent à l'utilisateur de visualiser ses données en utilisant différents types de diagrammes et de tableaux. Ils permettent aussi d'explorer les données à l'aide de différents opérateurs tels que le forage vers le bas (*drill-down, descendre dans la hiérarchie de la dimension*), forage vers le haut (*drill-up, roll-up, remonter dans la hiérarchie de la dimension*), le forage latéral (*drill-across, permet de passer d'un membre de dimension à un autre*) et le pivotage (*pivot, interchanger deux dimensions pour modifier le contenu des axes des diagrammes ou des tableaux visualisés*) (Rivest et al., 2004).

2.2.1.2 Types d'architectures OLAP

Il existe trois principaux types d'architectures OLAP, à savoir : (i) ROLAP (Relational Online Analytical Processing ou OLAP Relationnel), (ii) MOLAP (Multidimensional Online Analytical Processing ou OLAP Multidimensionnel) et (iii) HOLAP (Hybrid Online Analytical Processing ou OLAP Hybride).

ROLAP :

Dans une architecture ROLAP (Figure 10), les données sont stockées dans une base de données relationnelles. Le serveur ROLAP extrait les données de la base via des requêtes SQL et interprète ses données selon une vue multidimensionnelle avant de les présenter au module client.

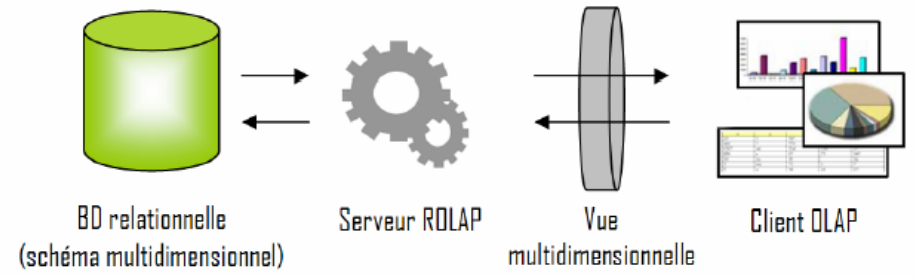


FIGURE 10 – Architecture système ROLAP source : (Bernier and Bédard, 2006)

MOLAP :

À l'inverse, dans une architecture MOLAP (Figure 11), les données stockées dans une base de données multidimensionnelles. Le serveur MOLAP a pour rôle d'extraire les données de la base de données multidimensionnelles avant de les présenter au module client (client OLAP).

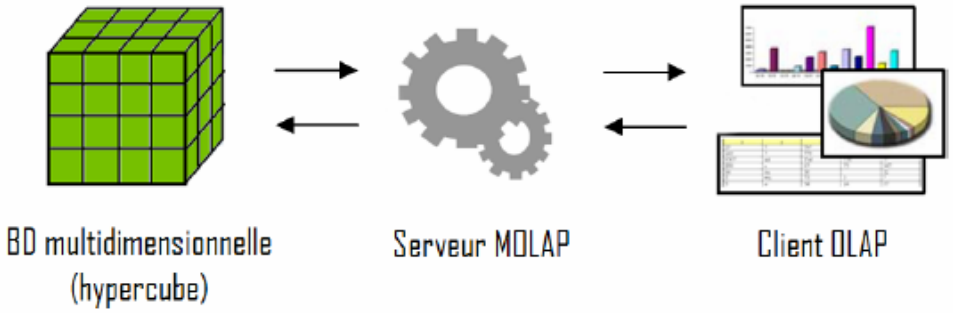


FIGURE 11 – Architecture système MOLAP souce : (Bernier and Bédard, 2006)

HOLAP :

Par contre, une architecture HOLAP (Figure 12) est la combinaison des architectures ROLAP et MOLAP. Les données sont stockées dans une base de données relationnelle et les données agrégées sont stockées dans la base de données multidimensionnelle. Le serveur HOLAP accède aux données stockées dans les bases de données relationnelle et multidimensionnelle avant de les présenter au client selon une vue multidimensionnelle.

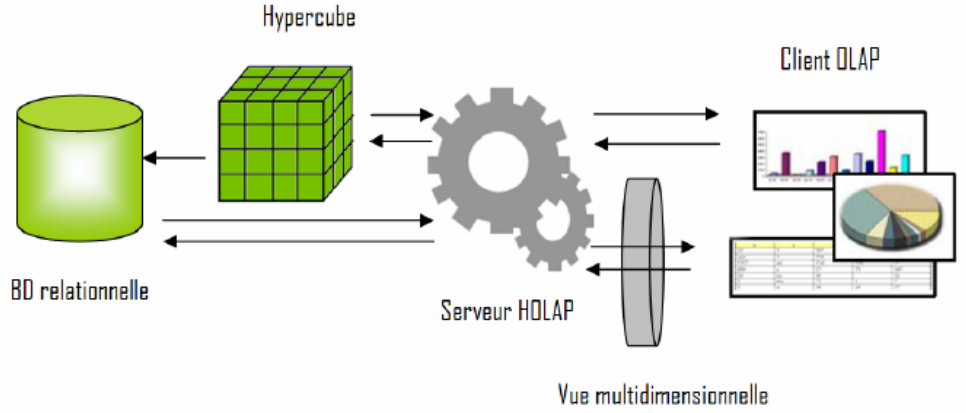


FIGURE 12 – Architecture système HOLAP source : (Bernier and Bédard, 2006)

2.2.2 Caractéristiques des systèmes OLAP

En 1993, E.F Codd a défini douze caractéristiques que doit respecter un système OLAP. Ces douze règles sont énoncées et décrites ci-après selon (Caron, 1998) :

- Vue conceptuelle multidimensionnelle des données : Cette règle part du principe que les analystes conçoivent les données relatives à leur entreprise de façon multi-dimensionnelle.
- Transparence : l'interface globale à l'usager devrait être cohérent. Cet usager pourra faire appel ou non aux fonctions OLAP pour ces besoins d'analyse.
- Accessibilité : le produit OLAP devrait être responsable d'aller chercher les données nécessaires à l'analyste de façon transparente, qu'elles se trouvent sur d'anciens systèmes, dans des fichiers de l'entreprise ou ailleurs, de façon à toujours lui présenter une image uniforme, c'est-à-dire au même format.
- Performance stable : Si le système était victime d'une dégradation de performance (temps de réponses prolongés) avec l'accroissement du nombre de dimensions, ou de niveaux d'agrégation impliqués dans une analyse, l'usager aurait à développer des méthodes de travail destinées à pallier cette dégradation, et aurait du même coup à dévier du processus d'analyse qu'il désirait mener à l'origine.
- Architecture client-serveur : Le produit OLAP doit pouvoir effectuer les opérations d'intégration nécessaires au respect de la transparence dans le cas où l'accès aux données se fait à partir de plusieurs sources hétérogènes.
- Dimensionnalité générique : Toutes les dimensions doivent être équivalentes, tant au niveau de leur structure (agrégations) que de leurs fonctionnalités. Ainsi, si certaines opérations sont souhaitables pour certaines dimensions, elles devront être disponibles pour toutes les dimensions.
- Gestion dynamique des cellules sans données : Le produit OLAP devrait être apte à gérer automatiquement les cellules (valeurs) pour lesquelles les données ne sont pas disponibles, impossibles à avoir ou inutiles, ainsi qu'à adapter son mode de

gestion des cellules vides selon le degré de dissémination de ces dernières.
— Support simultané de plusieurs utilisateurs : L'accès simultané au produit OLAP de la part de plusieurs utilisateurs doit être supporté, tout en préservant l'intégrité des données
— Liberté totale dans les opérations entre les dimensions : L'usager doit pouvoir, sans aucun sentiment de restriction, effectuer des opérations autant sur une dimension (forage, calcul de nouvelles valeurs, etc.) que sur plusieurs (analyses croisées, etc.)
— Manipulation intuitive des données : Le dialogue entre l'usager et son application doit se faire de la façon la plus intuitive possible, par des actions directes sur les valeurs ou les dimensions.
— Souplesse de création de rapports : Le produit doit permettre la création de rapports qui peuvent présenter une partie ou la totalité des données, dans une liberté de forme et une facilité de création nécessaires à la bonne présentation du phénomène souhaité.
— Nombre illimité de dimensions et de niveaux d'agrégation : L'usager doit pouvoir créer autant de dimensions qu'il désire avec autant de niveaux d'agrégation que souhaité.

2.2.3 Limites des systèmes OLAP

Selon (Bernier, 2002), « les outils OLAP qui sont actuellement sur le marché n'exploitent pleinement que la dimension descriptive des données. En effet, même si les données peuvent comprendre une certaine composante spatiale (adresse, code postal, etc.), elles sont analysées uniquement de façon descriptive (tableaux ou graphiques statistiques) et lorsqu'elles le font de façon cartographique, le résultat est très limité. Une analyse purement nominale des données spatiales est donc une limitation des outils OLAP actuels à supporter le processus d'analyse spatio-temporelle ». Les outils OLAP classiques sont limités quant à la présentation des données spatiales (Dubé, 2008). Voilà pourquoi, sans visualisation et navigation cartographiques, ces outils présentent d'importantes limitations pour l'analyse de phénomènes géographiques et spatio-temporels comme on en rencontre en environnement, en foresterie, en agriculture, en urbanisme, en sécurité, transport, etc (Caron, 1998). Tout comme dans les domaines cités par (Caron, 1998), en épidémiologie, la dimension spatiale est très importante. Car, elle permet de mieux comprendre la distribution spatiale d'une épidémie (par exemple la tuberculose, la grippe, la COVID-19). Cela permet de cibler les actions de riposte dans les zones auxquelles l'on suppose que l'apparition de la maladie est la plus probable. C'est pourquoi, l'utilisation d'outils OLAP classiques pour la compréhension des phénomènes de santé comme la tuberculose, la grippe, la COVID-19, etc. présente une limite. Car la dimension spatiale n'est pas prise en compte dans ce type de système.

2.2.4 Modélisation multidimensionnelle d'un entrepôt de données

Le modèle OLAP ou multidimensionnel organise les données en fonction des dimensions et des faits. Les dimensions représentent les axes d'analyse et sont organisées en hiérarchies. Une hiérarchie est composée de niveaux qui définissent les granularités d'analyse. Les faits sont les sujets d'analyse et sont décrits par des mesures (Bimonte, 2019). En général, les mesures sont des valeurs numériques qui sont agrégées à l'aide de fonctions d'agrégation numérique sur les niveaux de dimensions. Les données stockées sont ensuite explorées à l'aide des opérateurs OLAP, qui permettent de naviguer dans les hiérarchies (Roll-Up et Drill-Down), et de sélectionner un sous-ensemble de données de l'Entrepôt de Données (Slice and Dice) (Bimonte, 2019).

2.2.4.1 Modèles conceptuels d'un entrepôt de données

Au niveau conceptuel, nous distinguons trois schémas possibles pour modéliser un entrepôt de données : le schéma en étoile, le schéma en flocon et le schéma en constellation. Un schéma en étoile est constitué d'une table de fait centrale et des dimensions ou tables dimensionnelles (figure 13).

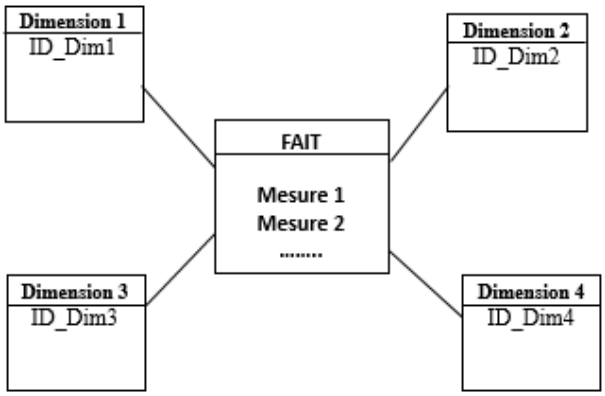


FIGURE 13 – Schéma en étoile

Le schéma en flocon (Figure 14) est un type de schéma en étoile qui inclut les formes hiérarchiques des dimensions ou tables dimensionnelles.

Le schéma en constellation (Figure 15) est une fusion de plusieurs schémas en étoile qui utilisent des dimensions communes.

2.2.4.2 Analyse comparative des schémas en étoile et en flocon de neige

Au Tableau 2.1, nous réalisons une analyse comparative des deux principaux modèles conceptuels d'un entrepôt de données : en étoile et en flocon de neige. Nous avons retenu trois critères (schéma du modèle, requêtes et temps d'exécution des requêtes) d'analyse

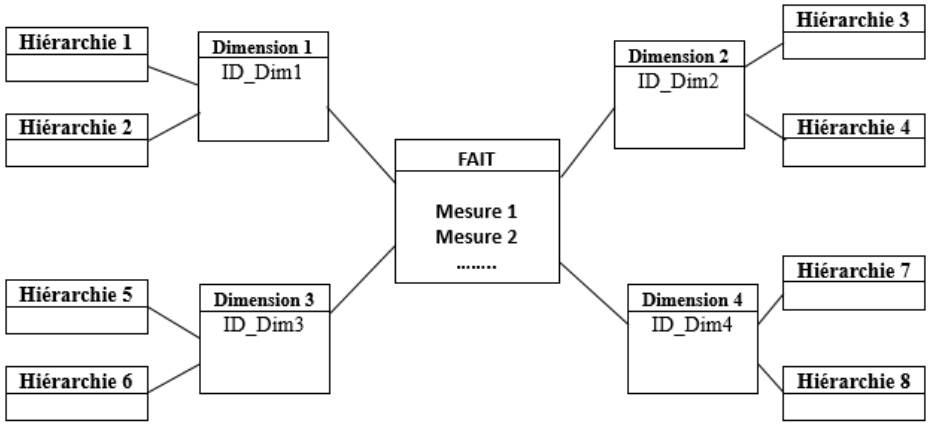


FIGURE 14 – Schéma en flocon

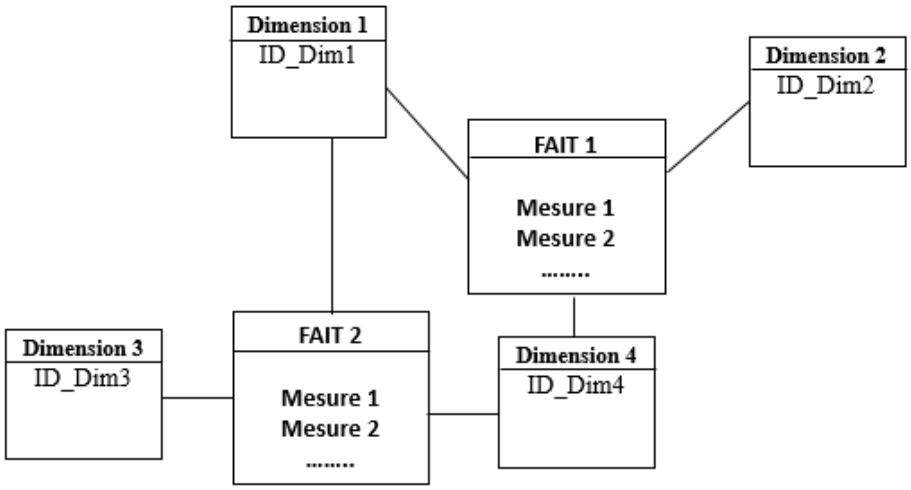


FIGURE 15 – Schéma en constellation

qui seront utiles pour la conception de l'entrepôt de données de la tuberculose et sa mise en œuvre.

TABLE 2.1 – Comparaison modèle en étoile et en flocon

Critères	**Modèle en étoile**	**Modèle en flocon**
Schéma	Table de fait et des tables de dimensions	Table de fait, tables de dimensions et tables de hiérarchies
Requêtes	Non complexe à cause du très faible nombre de jointures	Complexe à cause du grand nombre de jointures
Temps exécution requêtes	Faible	Long à cause des jointures

Les critères d'analyse montrent que le modèle conceptuel en étoile présente plus d'avantages au niveau des critères *requêtes* et *temps d'exécution des requêtes*. En effet, il y a un très faible nombre de jointures, ce qui rend le temps d'exécution des requêtes faibles. Aussi, le modèle conceptuel est simple, une table de fait au centre avec des dimensions autour.

2.2.5 Système d'information Géographique (SIG)

2.2.5.1 Définition

Dans la littérature, il existe un certain nombre de définitions des SIG. Parmi ces définitions, nous pouvons citer celle de (Joliveau, 1996) qui a décrit un SIG comme l'ensemble des structures, de méthodes, outils et données constitués pour rendre compte de phénomènes localisés dans un espace spécifique et faciliter les décisions à prendre sur cet espace. Pour (Chang, 2016), un SIG est un système informatique permettant de capturer, stocker, interroger, analyser et afficher des données géospatiales. Récemment, les SIG ont été définis comme un processus en plusieurs étapes : (i) acquisition des données, (ii) intégration en format SIG (saisie, géoréférencement), (iii) structuration, construction d'une base de données, (iv) interrogations, traitements, analyses, (v) résultats et (vi) communication (Denis, 2021).

Bien qu'il existe plusieurs définitions d'un SIG, toutes ces définitions ont néanmoins en commun le fait de reposer sur ses différentes fonctions.

2.2.5.2 Architecture d'un SIG

Un SIG est constitué d'un ensemble de composants (Figure 16). Ces composants sont reliés entre eux de manière hiérarchique. L'interface homme-machine définit la manière dont le système est exploité et contrôlé. À un niveau intermédiaire, un SIG doit disposer de mécanismes de traitement des données spatiales (entrée, édition, analyse, visualisation et sortie). À l'intérieur du système, une base de données géographiques stocke et extrait les données spatiales (Câmara et al., 2004).

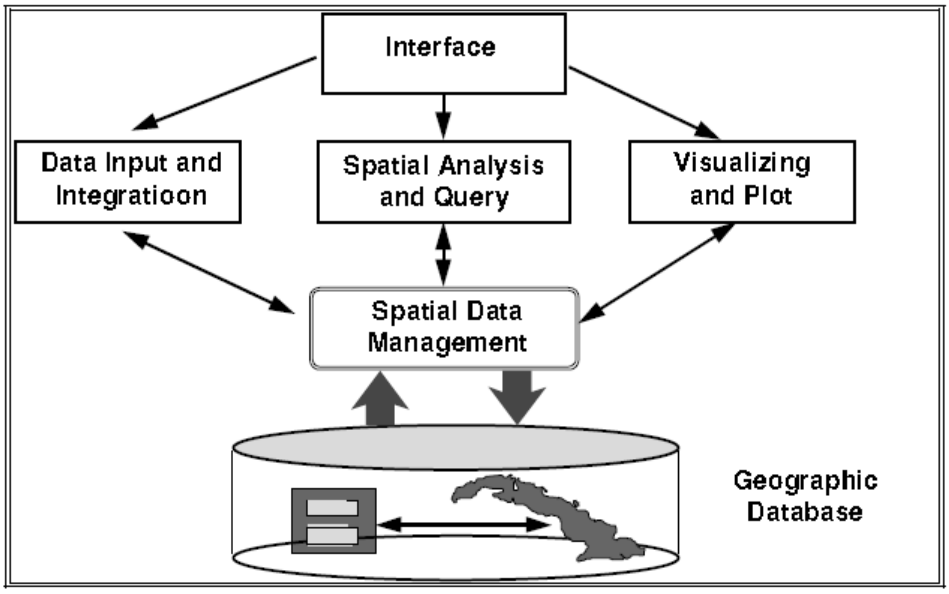


FIGURE 16 – L'architecture des systèmes d'information géographique (Câmara et al., 2004)

2.2.5.3 Fonctionnalités d'un SIG

Un SIG possède cinq principales fonctionnalités (les 5 A) :

— Abstraction : définition du modèle conceptuel de données. Ce qui se traduit par le choix des données à prendre en compte dans le SIG, leur définition et leur structuration. Des classes d'objets sont définies (par exemple arrondissement, quartier), avec leurs attributs (par exemple cordonnées géographiques, nom arrondissement, nom quartier), mais aussi avec leur relation (par exemple un arrondissement est composé de quartiers).

— Acquisition : C'est l'étape de collecte des données (géographiques et descriptives) qui vont alimenter le SIG ;

— Archivage : Il s'agit du stockage des données. Elles peuvent être stockées dans un répertoire (fichiers forme ou Shapefile, CSV, etc.) ou dans un système de gestion de base de données (SGBD) spatial (par exemple PostGIS, Oracle Spatial) ;

— Analyse : À partir des données stockées dans le SIG, il s'agit de répondre à un certain nombre de questions en réalisant des analyses spatiales ;

— Affichage : Il s'agit de la présentation des résultats d'analyses sous forme de cartes thématiques.

2.2.5.4 Domaines d'application des SIG

Depuis de nombreuses années, les SIG ont été appliqués à des domaines variés :

— La gestion des risques naturels (Chatelain et al., 1995; Pareschi et al., 2000; Lacambre, 2001; Gourmelon, 2003) ;

— L'agriculture (Wilson, 1999; Passouant et al., 2000; Ariaux, 2000; Bill et al., 2011; Kedowide Mevo Guezo, 2011) ;
— Les études des statistiques (Leroi et al., 2001; Bonin et al., 2001; Mancini et al., 2010) ;
— La santé (Lang, 2000; McLafferty, 2003; Lacoste, 2006; Taravat, 2009)

Dans le domaine de la surveillance épidémiologique des maladies, la dimension spatiale est très importante. En effet, l'exemple « historique » de l'utilisation de la dimension spatiale en épidémiologie est l'analyse de l'épidémie de choléra à Londres au XIXe siècle par John Snow (Snow 1855). Le simple fait de reporter les cas sur un plan de quartier permit à celui-ci : (i) d'identifier le lieu de la contamination, (ii) d'améliorer les connaissances sur le mode de transmission de la maladie et, surtout, (iii) d'intervenir et d'arrêter l'épidémie (Tran et al., 2009).

2.2.5.5 Limites des SIG

Les SIG conventionnels présentent des limites. Parmi celles-ci, nous pouvons citer (McHugh et al., 2007) : (i) les interfaces des SIG conventionnels sont trop complexes pour des non-experts et (ii) le niveau de complexité des SIG conventionnels est beaucoup trop élevé pour permettre aux utilisateurs de réaliser à la volée des croisements de données, des analyses spatio-temporelles, d'explorer les données selon différentes perspectives ou selon différents niveaux de détails, et ce, sans devoir maîtriser un langage d'interrogation. Enfin, les SIG sont bien adaptés pour l'analyse de données spatio-temporelle, par contre, ils ne sont pas capables d'effectuer une analyse multidimensionnelle (Younsi, 2016).

2.2.6 Système d'information géographique décisionnel

Les SIG n'étant pas adaptés pour effectuer des analyses multidimensionnelles, une nouvelle approche combinant les fonctionnalités d'un SIG et d'un système décisionnel (OLAP), le SOLAP (Spatial On-Line Analytical Processing) a été proposée par l'équipe de recherche en bases de données spatiales du Centre de recherche en géomatique du Canada, dirigée par le Dr Yvan Bédard. Par la suite, de nombreux travaux de recherche (Chaker et al., 2009; Bimonte et al., 2014, 2016b) utilisant cette approche ont été réalisés.

2.2.6.1 Définition

Un système d'information géographique décisionnel ou SOLAP est « une plate-forme visuelle spécialement conçue pour supporter l'analyse et l'exploration spatio-temporelles rapides et faciles des données multidimensionnelles à l'aide d'affichages cartographiques aussi bien qu'à l'aide de tableaux et diagrammes statistiques » (Andrienko and Andrienko, 1999). Pour (Bédard, 2004), c'est « un type de logiciel qui permet la navigation rapide et

facile dans les bases de données spatiales et qui offre plusieurs modes d'affichage synchronisés ou non : cartes, tableaux et diagrammes ».

2.2.6.2 Architecture SOLAP

De façon générale, les systèmes SOLAP sont implémentés suivant une architecture composée de (Figure 17) : un ETL (Extraction-Transformation-chargement) spatial, un entrepôt de données spatiales (EDS), un serveur SOLAP et un client SOLAP (Bimonte et al., 2016a). En effet, les données issues des sources hétérogènes doivent préalablement subir un prétraitement via l'ETL avant d'être chargées dans l'entrepôt de données spatiales.

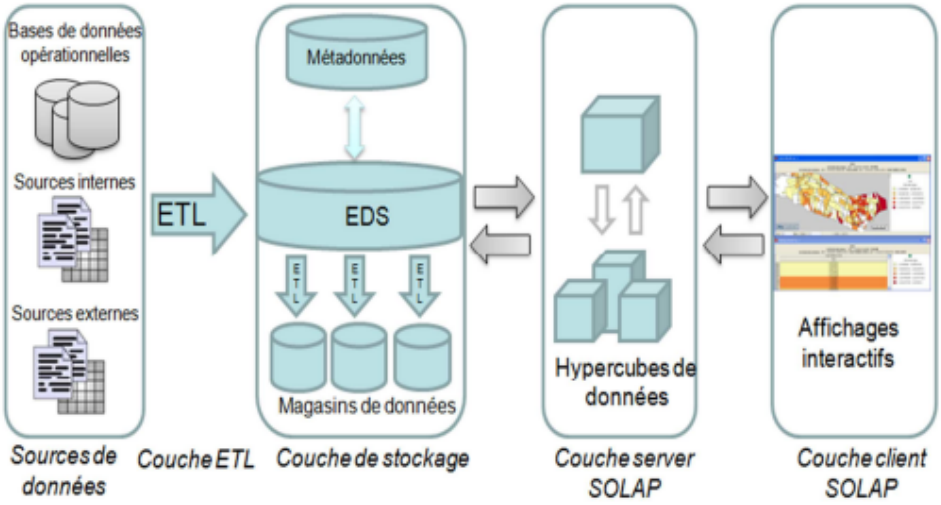


FIGURE 17 – Architecture SOLAP (Bimonte et al., 2016a)

L'entrepôt de données spatiales (EDS) est conçu en utilisant un système de gestion de bases de données relationnel (SGBDR) spatial (par exemple PostGIS). Ce niveau est responsable du stockage des données multidimensionnelles, alphanumériques et spatiales (Bimonte et al., 2016b).

Le serveur SOLAP est responsable de la mise en œuvre des opérateurs SOLAP pour calculer et naviguer au sein des cubes de données spatiales (Bimonte et al., 2016b). Ces cubes de données sont représentés par un schéma SOLAP qui définit une correspondance entre les concepts de cube de données spatiales (dimensions, hiérarchies, mesures, etc.) et le schéma relationnel (tables et colonnes utilisées pour représenter les faits et les dimensions) (Bimonte et al., 2016b).

Le client SOLAP est composé d'un client OLAP et un client SIG (Bimonte et al., 2016b). Le client OLAP a pour rôle de rendre l'information multidimensionnelle « visible », en d'autres termes, de permettre de découvrir des connaissances grâce à la seule visualisation et interaction avec les données (Bimonte, 2007). Le client SIG est le responsable de la visualisation cartographique (Bimonte et al., 2016b) des données.

2.2.6.3 Approche de conception d'un SOLAP

Pour concevoir une application SOLAP, trois familles de solutions sont disponibles (Bédard et al., 2005) :

- Solutions OLAP dominant : ils procurent toutes les fonctionnalités d'un outil OLAP, il est donc implicite qu'un tel outil utilise les capacités d'un serveur OLAP. Par contre, cette solution n'intégrera qu'un très faible sous-ensemble des fonctions des systèmes d'information géographique (SIG), généralement les fonctions d'affichage, de navigation cartographique (par exemple déplacement et changement d'échelle) et de sélection d'éléments géométriques. .
- Solutions SIG dominant : Ces solutions offrent toutes les fonctionnalités d'un outil SIG, mais seulement un sous-ensemble des fonctionnalités de l'outil OLAP. Cette solution couple une base de données relationnelle simulant un serveur OLAP à un logiciel SIG ou à un outil de visualisation de données spatiales. .
- Solutions hybrides ou intégrées : Ces solutions intègrent les fonctionnalités d'un outil OLAP et d'un SIG. Ce type de solution est utile lorsque l'application doit s'intégrer dans un environnement géomatique à fort flux de données (par exemple pour les mises à jour cartographiques) ou nécessite l'utilisation de fonctions spécifiques au SIG, comme les opérateurs d'analyses spatiaux (par exemple l'opérateur de forage spatial, qui permet à l'usager de naviguer d'un niveau général à un niveau plus détaillé à l'intérieur d'une dimension géographique : cartographier les régions sous-jacentes composant un pays).

2.2.6.4 Outils d'analyse et de visualisation de données en informatique décisionnelle

Les outils d'analyse et de visualisation des données peuvent être un moyen efficace à utiliser dans de nombreux domaines (par exemple finance, commerciale, santé). Car, à partir des données et des tableaux de bord produits à partir ces outils, on peut faire émerger des tendances, des comparaisons ou encore des recommandations qui sont utiles pour l'aide à la décision. Particulièrement dans le domaine de la santé, ce type d'outil peut aider à améliorer les décisions des autorités sanitaires dans la riposte contre les maladies. En informatique décisionnelle, il existe de nombreux outils permettant de réaliser des analyses multidimensionnelles de données et de visualiser les résultats d'analyses sous formes de tableaux croisés, graphiques, cartes.

Parmi ces outils, le top 10 en 2022 [5] était composé respectivement de : (1) Google data studio, (2) Tableau, (3) Power BI Pro, (4) Looker, (5) Qlik Sense, (6) Mode, (7) Chartio, (8) DOMO, (9) IBM Cognos Analytics, (10) Sisense. En nous appuyant sur certains travaux comme ceux de (Gowthami and Kumar, 2017; Barreto, 2019; Lousa et al., 2019),

5. https ://www.boryl.fr/blog/top-10-des-outils-de-business-intelligence/

nous avons réalisé une analyse comparative des outils Tableau et Power BI qui ont retenu notre attention. Nous avons opté pour ces deux outils, car, Google data studio qui est l'outil BI le plus utilisé avant Tableau et Power BI est basé exclusivement sur le cloud. Or, la solution que nous proposons doit permettre aux décideurs et aux experts de la santé publique de pouvoir visualiser les tableaux de bord d'indicateurs sans forcément avoir besoin d'un accès à Internet. Ce qui est possible avec Tableau et Power BI.

Nous avons retenu les critères d'analyse suivants (Tableau 2.2) : interface utilisateur, faciliter de prise en main par les futurs utilisateurs, sources de données, ETL, connexion OLAP, analyses géospatiales et visualisation des données.

TABLE 2.2 – Analyse comparative des outils d'analyse et de visualisation des données

Critère	**Power BI**	**Tableau**
Interface utilisateur	Bonne	Meilleure pour les non-informaticiens
Faciliter de prise en main par les futurs utilisateurs	Connaissance d'Excel	Ne nécessite aucune compétence technique ou de programmation pour être utilisé par les non-informaticiens
Sources de donnée	Accès limité à d'autres bases de données et serveurs (portefeuille de Microsoft), par rapport à Tableau	Accès à de nombreuses sources de données et à de nombreux serveurs
Outil ETL	Dataflow	Tableau Prep
Connexion OLAP	se connecte aux cubes OLAP via des serveurs SQL pour une analyse multidimensionnelle	se connecte à OLAP en extrayant les mesures du cube au niveau le plus profond
Analyse géospatiale	Oui	Mieux classé que Power BI pour les capacités de visualisation géospatiale et visualisation cartographique
Visualisation des données	Lignes de tendance, graphiques, cartes, tableaux de bord, etc.	Graphiques, cartes, tableaux multiples, tableaux de bord, etc.

Par rapport aux objectifs de notre travail de thèse et des besoins des futurs utilisateurs qui sont pour la plupart des non-informaticiens, nous retenons l'outil Tableau qui présente plus d'avantages que Power BI. Parmi ces avantages, nous pouvons citer : (i) meilleure interface utilisateur pour les non-informaticiens, (ii) ne nécessite aucune compétence technique, (iii) accès à des nombreuses sources de données, (iv) tout comme Power BI intègre un outil ETL (Tableau Prep) et (v) mieux classé que Power BI dans la visualisation géospatiale et la visualisation cartographique.

Il faut néanmoins relever que certains types d'analyses spatiales, par exemple la détection d'agrégats spatiaux en utilisant les méthodes comme l'indice des proches voisins (distance entre cas), foyers d'agrégation (présence de foyers de cas), etc, ne peuvent pas être réalisées directement avec ces outils BI. D'autres outils ou logiciels doivent être utilisés pour réaliser ce type d'analyse (par exemple RStudio, GeoDa).

2.2.6.5 SOLAP en épidémiologie

De nombreux travaux utilisant la technologie SOLAP ont été menés dans divers domaines (économie, agriculture, transport, santé publique, etc.). Nous faisons un focus sur les SOLAP en santé publique qui ont retenu notre attention. Parmi ceux-ci, nous pouvons citer celui proposé par (Proulx et al., 2002). Les auteurs ont utilisé une approche SIG dominant pour développer une application dans le domaine de la santé publique en exploitant la technologie SOLAP pour découvrir la relation entre la santé respiratoire et l'environnement (par exemple l'incidence des maladies respiratoires en fonction de la qualité de l'air). Une base de données a été conçue sous Access pour le support multidimensionnel, SoftMap a été utilisé pour la cartographie et Visual Basic comme langage de développement. En outre, le modèle conceptuel de données repose sur une architecture multidimensionnelle. En effet, dans l'outil proposé, on peut changer de dimension. On peut également réaliser des analyses temporelles et multi-échelles des données spatiales via les fonctionnalités de navigation spatiale (c.-à-d. forage spatial) sur les données.

Dans les travaux de (Scotch and Parmanto, 2005), l'outil SOVAT (Spatial OLAP Visualisation and Analysis Tool) a été conçu pour aider les chercheurs et les professionnels de la santé publique dans le processus de prise de décision. Les principales caractéristiques de cet outil sont les suivantes : (i) analyses statistique et spatiale, (ii) l'exploration détaillée des données et (iii) la visualisation des données au moyen de graphiques et de cartes.

Un système d'exploration de données automatisé qui permet aux décideurs de la santé publique d'accéder aux informations concernant les tumeurs cérébrales a été développé par (Santos et al., 2013). Ils ont construit un entrepôt de données et utilisé l'ontologie dans un processus de data mining automatisé. Il s'agit d'une ontologie [6] pour représenter le domaine des tumeurs cérébrales. Les auteurs ont utilisé un référentiel d'ontologies existant, le thésaurus NCI [7] en OWL. En outre, les principales activités de fouille de données sont : (i) l'association (consiste à identifier des profils de classes en fonction de leurs attributs et à déterminer à laquelle de ces classes prédéfinies appartient un nouvel élément. Par exemple, la plupart des patients qui reçoivent des ordonnances pour le médicament A reçoivent également des ordonnances pour le médicament B), (ii) la

6. L'ontologie médicale est l'étude de ce qui « est » en médecine et du processus de leur formation. Elle s'intéresse à la genèse des entités médicales : les maladies, les signes cliniques, les syndromes cliniques..., (www.encyclopedie.fr)

7. https ://evs.nci.nih.gov/evs-download/thesaurus-downloads

classification (Par exemple, étant donné des classes particulières de patients ayant des réponses différentes aux traitements médicaux, la classification est utilisée pour identifier la forme de traitement à laquelle un nouveau patient est le plus susceptible de répondre) et (iii) le regroupement (utilisé pour identifier des groupes d'éléments qui sont similaires. Par exemple, à partir d'un ensemble de données sur les patients, le regroupement peut être utilisé pour identifier des sous-groupes de patients ayant des schémas de traitement similaires).

Un outil SOLAP pour la surveillance et la prédiction du phénomène de la propagation de l'épidémie de grippe saisonnière a été développé par (Younsi et al., 2019). Cet outil est composé d'un entrepôt de données et d'une interface pour la visualisation (carte, graphique, ..) de l'évolution spatio-temporelle de la maladie.

Par contre, (Amina et al., 2011; Benabbou et al., 2014) ont utilisé une approche SOLAP pour la surveillance de la tuberculose.

Dans (Amina et al., 2011) une démarche décisionnelle pour la surveillance des maladies décomposée en trois phases a été proposée :

- Structuration : Il s'agit du traitement des données issues dans différentes sources avant leur stockage dans l'entrepôt de données multidimensionnelles ;
- Exécution : A cette phase, le serveur de base de données extrait les données et interprète les données selon une vue multidimensionnelle avant de les présenter au module client selon différents modes d'affichage ;
- Concrétisation : Cette phase s'intéresse à l'interprétation des résultats d'analyse et la découverte de la connaissance afin de faciliter la prise de la décision pour les acteurs de la santé.

Dans leur étude, ils se sont focalisés sur les deux premières phases. Un exemple d'application de ce processus à l'épidémie de tuberculose a été proposé. Ainsi, ils ont conçu un entrepôt de données spatiales de la tuberculose. Le fait étudié est l'incidence de la tuberculose. Il a pour dimensions le temps (année, trimestre), spatiales (commune, secteur socio-sanitaire, région), sexe (homme ou femme), âge (âge, groupe âge), type de maladie (pulmonaire ou extra-pulmonaire). Il faut relever que le cube de données spatiales de la tuberculose proposé ne prend en compte la dimension *résultat thérapeutique* qui est pourtant très essentiel pour bien mesurer l'incidence de la tuberculose au sein de population.

Les travaux de (Benabbou et al., 2014) ont consisté à la conception d'un outil d'aide à la décision basé sur l'approche SOLAP pour la surveillance de la tuberculose. Le sujet d'analyse ou fait était le "Traitement". Il avait pour axes d'analyse ou dimensions "Résultat thérapeutique", "Temps", "Espace", "Type de malade" et "Résultat culture". L'outil repose sur une architecture à trois niveaux :

(i) *niveau visualisation* : il s'agit de l'interface utilisateur qui permet de présenter l'information spatiale et textuelle aux décideurs qui assure trois modes d'affichage des données sur une carte interactive : données brutes, pourcentage et camembert ;

(ii) *niveau logique* : il s'agit du gestionnaire de requêtes responsable de la résolution des requêtes OLAP demandées par le décideur et de l'alimentation de l'interface utilisateur ;

(iii) *niveau données* : ce niveau est chargé du stockage et de la recherche d'information dans les entrepôts de données descriptives et la base de données des frontières administratives géographiques de la zone d'étude.

L'interface utilisateur a été réalisée à l'aide de la bibliothèque java GeoTool, la communication entre la base de données et le gestionnaire de requête se fait à travers le moteur OLAP Mondrian et les données sont stockées dans une base de données relationnelle sous MySQL.

Nous avons réalisé une comparaison des solutions SOLAP pour la surveillance de maladies proposées dans l'état de l'art (tableau 2.3). Les critères de comparaison sont les suivants : (i) représentation cartographique, (ii) tableaux et diagrammes, (iii) analyse multidimensionnelle, (iv) entrepôt de données spatiales et (v) surveillance tuberculose.

TABLE 2.3 – Analyse comparative des SOLAP de surveillance de maladies

Critère	(Proulx et al., 2002)	(Scotch and Parmanto, 2005)	(Santos et al., 2013)	(Benabbou et al., 2014)	(Younsi et al., 2019)
Représentation cartographique	Oui	Oui	Non	Oui	Oui
Tableaux et diagrammes	Oui	Oui	Non	Oui	Oui
Analyses multidimensionnelles	Oui	Oui	Oui	Oui	Oui
Entrepôt de données spatiales	Oui	Oui	Non	Non	Oui
Surveillance tuberculose	Non	Non	Non	Oui	Non

2.2.7 Conclusion

Dans la littérature, nous avons pu retrouver qu'une seule proposition d'outil SOLAP appliqué à la surveillance épidémiologique de la tuberculose. Dans cet outil, nous avons relevé que : (i) la base de données géographique se résumait à trois fichiers formes (Shape files) des frontières géographiques de la région d'étude et (ii) au niveau du modèle multidimensionnel, à la dimension "résultat thérapeutique", les patients perdus de vue sous traitement antituberculeux ne sont pas pris en compte. Or, le suivi des patients perdus de vue est très important pour les autorités sanitaires et les experts de la santé afin d'éviter la propagation de la maladie. D'autres dimensions également très importantes (statut VIH, statut professionnel) pour la surveillance de l'infection tuberculeuse n'ont pas été

pas prises en compte dans le modèle conceptuel. Aussi, ce SOLAP n'intégrait pas de plate-forme d'enregistrement des patients tuberculeux. Ce qui ne permettait pas la mise à jour des données stockées dans l'entrepôt. Entraînant ainsi un problème de fraîcheur de ces données.

Dans le cadre de ce travail, nous proposons une nouvelle approche de SOLAP de la tuberculose. En s'appuyant sur les besoins d'analyse recueillis auprès des autorités sanitaires et les experts de la santé publique, notre modèle conceptuel multidimensionnel intègre toutes les dimensions (âge, genre, statut professionnel, statut VIH, type de tuberculose, type patient, résultat thérapeutique, temps, géographie) nécessaires à une meilleure compréhension de l'épidémie de la tuberculose. Nous couplons un hypercube de données spatiales avec un outil d'analyse et de visualisation de données multidimensionnelles. Aussi, afin de permettre à chaque intervenant de la lutte antituberculeuse de disposer d'un outil adapté à ses besoins d'analyse (par exemple, suivi de l'évolution spatio-temporelle des cas, de la létalité, de l'incidence, simulation inter-individu de la TB, etc.), nous proposons une approche hybride de système d'information décisionnel qui couple le système SOLAP avec un système multi-agents (SMA).

Les résultats d'analyses multidimensionnelles produits par un SOLAP pour la surveillance épidémiologique doivent pouvoir être communiqués facilement aux experts de la santé publique qui sont des décideurs pour une riposte efficace. C'est dans cette direction que se positionnent les travaux concernant les processus de narration de données et d'intelligence épidémique, comme décrit à la section 2.3.

2.3 Processus de narration de données et d'intelligence épidémique

Dans cette section, nous passons en revue les travaux réalisés sur les processus de narrations de données classiques (sous-section 2.3.1) et d'intelligence épidémique (sous-section 2.3.2). L'objectif est de comprendre comment ont été construits ces différents processus afin de proposer un processus qui tienne compte à la fois des bonnes pratiques et des méthodologies en narration de données classiques et des spécificités d'un processus d'intelligence épidémique.

2.3.1 Processus de narration de données

La narration de données reçoit actuellement un intérêt croissant dans plusieurs domaines comme le journalisme des données, les affaires, le e-gouvernement, la science des données[8] (El Outa et al., 2021) et la santé (Rebman et al., 2017). Car, elle permet de

8. La science des données est un domaine interdisciplinaire qui utilise des méthodes, des processus, des algorithmes et des systèmes scientifiques pour extraire des connaissances et des idées à partir de

transmettre efficacement des informations importantes afin d'aider les décideurs à prendre de meilleures décisions.

Dans cette sous-section, au paragraphe 2.3.1.1, les concepts fondamentaux sur la narration de données sont présentés. Ensuite, les paragraphes 2.3.1.2 et 2.3.1.3 présentent respectivement les travaux connexes sur les processus de narration de données et les narrations de données qui ont été proposées dans le domaine de la santé publique. Nous terminons par une conclusion.

2.3.1.1 Concepts de la narration de données

De nombreuses définitions ont été données au concept de "narration de données". Elle a d'abord été décrite par (Carpendale et al., 2016) comme l'activité de production de narrations, étayée par des faits, extraits de l'analyse de données, à l'aide de visualisations potentiellement interactives. Tout récemment, elle a été définie comme une composition structurée de messages qui (a) transmettent des trouvailles sur les données et (b) les communiquent généralement par des moyens visuels afin de faciliter leur réception par un public cible (El Outa et al., 2021).

Pour formaliser les principaux concepts de la narration de données et leurs relations, (El Outa et al., 2020) ont proposé un modèle conceptuel en quatre couches (Figure 18) :

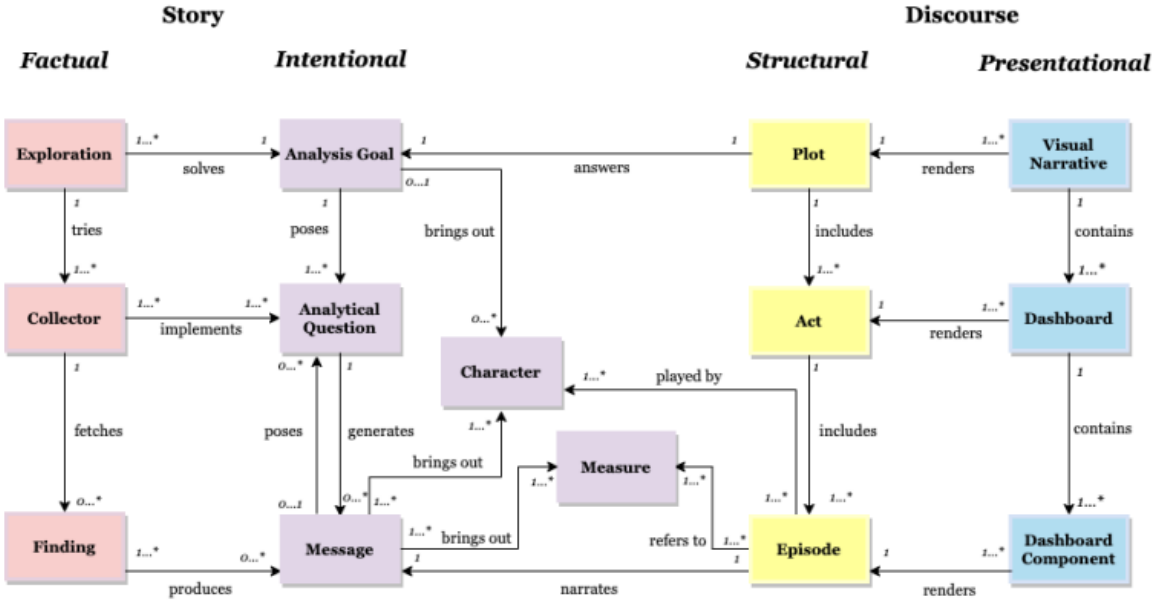


FIGURE 18 – Modèle conception pour la production d'une narration de données (El Outa et al., 2020)

— Couche factuelle : elle a trait à *l'exploration des sources de données* (par exemple dans ce travail, les données socio-demographiques et cliniques des patients tuber-

nombreuses données structurées ou non. Elle est souvent associée aux données massives et à l'analyse des données (wikipedia).

culeux, les données de la population et les données géographiques) à travers un ensemble de collecteurs qui permettent de manipuler les données avec des outils variés. Les trouvailles (exemple : 74% de patients tuberculeux perdus de vue, 61% de patients tuberculeux hommes, 54% de patients tuberculeux provenant des quartiers précaires, etc.) issues des données explorées sont des candidates pour participer à l'histoire.

— Couche intentionnelle : cette couche correspond à la phase de *définition de l'objectif de la narration*. Elle modélise la substance de l'histoire, en identifiant les messages à communiquer. Ces messages répondent à des questions analytiques (par exemple : quelles est la situation épidémiologique de la maladie ? quelle est la distribution spatiale des cas ?) que se pose le producteur de la narration, selon son objectif d'analyse (par exemple : décrire le profil épidémiologique de COVID-19). Les questions analytiques sont implémentées par les collecteurs, leurs trouvailles permettant de produire des messages (par exemples : il n'y a pas de corrélation spatiale ou spatio- temporelle des cas TB, l'évolution spatiale et temporelle est corrélée avec la typologie des quartiers).

— Couche structurelle : elle concerne la structuration du discours. Elle vise à organiser les différents messages de l'intrigue en termes d'actes et épisodes. Par exemple, à la figure 19 qui présente la situation de la COVID-19 en France, l'intrigue est organisée en deux actes. L'acte 1 comprend deux épisodes qui racontent respectivement le nombre de morts à l'hôpital pour 100 000 habitants et l'évolution temporelle de la positivité des tests par classe d'âge. L'acte 2 comprend également deux épisodes qui racontent les personnes nouvellement hospitalisées et en réanimation ou en soins intensifs.

— Couche présentationnelle : elle modélise le rendu de la narration de données, c'est-à-dire la narration visuelle qui est communiquée à l'audience via des artefacts visuels : tableaux de bord et composants. Par exemple, la narration visuelle peut être publiée sous forme d'une page Web contenant deux tableaux de bord qui vont restituer les deux actes de l'intrigue (Situation de la COVID-19 en France). Les composants des tableaux de bord sont responsables du rendu des épisodes avec plusieurs artefacts visuels : cartes thermiques (épisodes de l'acte 1) et cartes qui présentent la distribution spatiale des personnes nouvellement hospitalisées et en réanimation ou soins intensifs (épisodes de l'acte 2).

Après la présentation des principaux concepts sur la narration de données, nous passons maintenant aux travaux qui ont porté sur les processus de narration de données.

2.3.1.2 Travaux connexes sur les processus de narration de données

La narration de données, c'est-à-dire l'élaboration d'un récit de données, est un processus complexe à la croisée de plusieurs domaines : traitement des données, analyse des

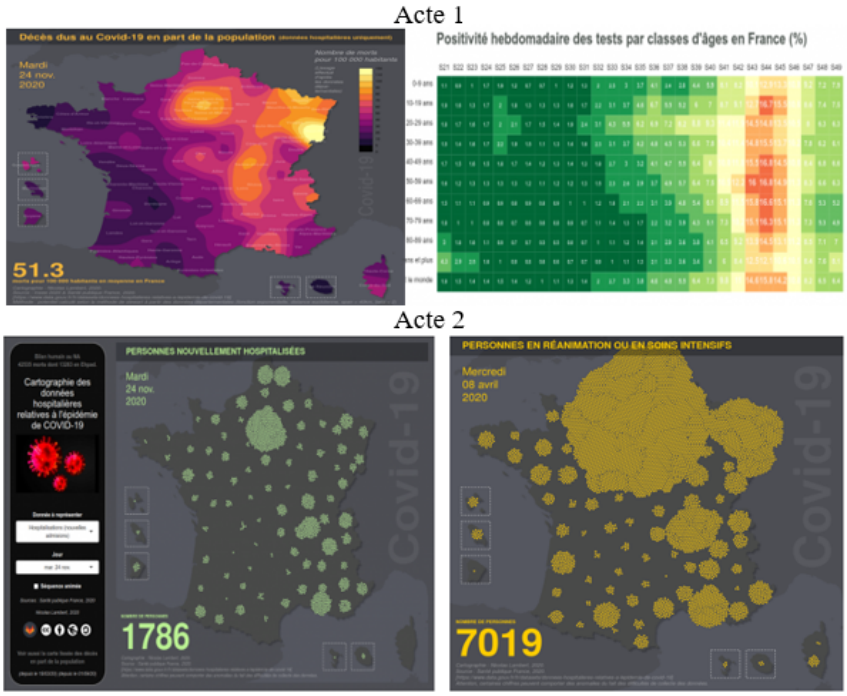


FIGURE 19 – Situation de la COVID-19 en France, composition d'images extraites de (Lambert, 2021)

données, visualisation des données, communication, entre autres. Cette section passe en revue les travaux sur les processus de narration de données.

D'abord, dans (Lee et al., 2015), le processus est établi en trois phases : (i) explorer les données, pour retrouver les résultats (faits saillants) parmi les données, (ii) créer l'histoire, pour transformer les faits-saillants en une séquence d'éléments narratifs et construire l'intrigue du récit, et (iii) raconter une histoire, pour rendre l'intrigue par des moyens visuels (par exemple, diaporama, vidéo et démonstration). Il faut relever que dans ce processus, pendant la phase de création de l'histoire, il est possible de revenir à la phase d'exploration des données pour recueillir d'autres faits saillants.

Ensuite, dans (Chen et al., 2018), le processus de narration proposé se décompose en deux principales phases : l'analyse des données, qui nécessite l'exploration de données complexes et leurs interrelations en utilisant des techniques d'interaction sophistiquées, et le rendu de la narration, qui concerne la transmission des informations intéressantes ou importantes extraites de l'analyse des données, présentées de manière simple et facilement compréhensible. Aussi, les auteurs ont proposé une phase intermédiaire, la synthèse des données, au cours de laquelle l'analyste rassemble et organise les éléments d'information à communiquer, facilitant ainsi la présentation des résultats de l'analyse visuelle dans un récit convaincant.

Enfin, plus récemment, dans (El Outa et al., 2022), un processus complet d'élaboration d'une narration de données a été proposé. Ce processus se décompose en quatre phases dans lesquelles un ensemble d'activités sont réalisées :

— Exploration : A cette phase, 5 activités sont réalisées (i) la collecte de données (la sélection des sources, l'extraction, l'intégration et le prétraitement des données), (ii) l'essai et la réutilisation de plusieurs collecteurs (c'est-à-dire des outils d'interrogation, de profilage et d'exploration) et (iii) l'essai de diverses visualisations (tableaux croisés, graphiques, cartes, etc.) pour la collecte des résultats (iv) la formulation des résultats, concernant le stockage des résultats et de leurs relations, et (v) la validation des résultats, qui se fait généralement par des tests statistiques.

— Réponse aux questions : A cette phase, 3 activités sont réalisées, (i) formuler des objectifs et des questions analytiques, (ii) tirer des messages, des résultats, et (iii) la validation des messages.

— Structuration : A cette phase, 4 activités sont réalisées, (i) déterminer le public cible, (ii) sélectionner éventuellement un sous-ensemble de messages pour ce public, et (iii) choisir une structure narrative appropriée. Ensuite, (iv) les messages sont mis en correspondance avec les actes et les épisodes.

— Présentation : A cette phase, 3 activités sont réalisées, (i) la définition du type de narration visuelle, (ii) la définition du mode d'interactivité et (iii) la conception de tableaux de bord pour transmettre les actes et les épisodes au public.

En somme, en comparant les processus de narration de données généraux que nous avons pu étudier (Chen et al., 2018; Lee et al., 2015; El Outa et al., 2022) (Figure 20), nous avons observé que chacun de ces processus est composé des phases exploration des données, structuration de la narration et présentation. Seul (El Outa et al., 2022) proposent une phase intermédiaire (réponse aux questions) avant la phase de structuration. Le processus de (El Outa et al., 2022) apparaît plus complet en termes de phases et d'activités réalisées à chaque phase par rapport ceux de (Chen et al., 2018; Lee et al., 2015).

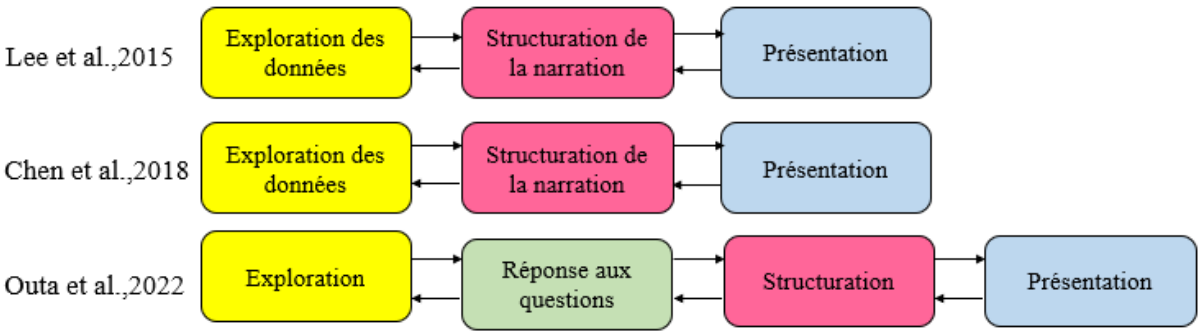


FIGURE 20 – Processus de narration de données généraux (Lee et al., 2015; Chen et al., 2018; El Outa et al., 2022)

Après l'étude des processus de narration de données, nous passons maintenant aux travaux qui ont porté sur la production de narrations de données dans le domaine de la santé publique. L'objectif est de comprendre les méthodes ou processus qui ont été

utilisées pour la production de ces narrations.

2.3.1.3 Travaux narration de données en santé publique

Il y a une tendance croissante à l'utilisation de la narration de données comme outil de recherche et d'intervention sur les questions de santé publique, en particulier celles qui comportent une forte composante de prévention des maladies (McCall et al., 2019). En effet, la narration de données était souvent utilisée pour éduquer les populations sur les pratiques de protection de la santé, pour plaider en faveur de l'amélioration des soins cliniques et pour encourager les efforts de lutte contre les maladies infectieuses (Tsui and Starecheski, 2018).

À titre d'exemples, (Larkey and Gonzalez, 2007) ont testé cette technique de communication pour promouvoir les messages de prévention et le dépistage du cancer colorectal chez les Latinos. Aussi, dans (Ezegbe et al., 2018), les auteurs ont mené une enquête pour déterminer l'efficacité d'une thérapie narrative [9] numérique rationnelle et émotionnelle sur la connaissance du VIH/sida et la perception des risques chez les écoliers de l'État d'Enugu, au Nigeria.

Dans (Njeru et al., 2015), les auteurs ont élaboré une intervention de narration numérique sur le diabète afin de sensibiliser les populations d'immigrants et de réfugiés ayant une maîtrise limitée de l'anglais.

Dans une autre étude menée dans des communautés rurales d'Alaska, (Cueva et al., 2015) ont cherché à savoir si les histoires numériques pouvaient influencer les sentiments des participants à l'égard du cancer, et si le visionnage des histoires numériques entraînait un changement (ou une intention de changement) dans le comportement de santé.

Ces études ont montré que la narration d'histoires peut augmenter les réponses positives des patients et augmenter leur niveau de connaissance sur ces maladies.

Dans (Botsis et al., 2020), un cadre théorique pour la construction d'une histoire visuelle en sciences biomédicales a été proposé. Le processus de construction comprend quatre phases : l'identification du public cible, l'évaluation de la littératie en santé [10] de l'audience, la conception de l'histoire visuelle et la production de l'histoire visuelle. Une application ce cadre théorique a donné lieu à la production de deux histoires visuelles, l'une sur la sécurité des vaccins et l'autre sur l'immunothérapie du cancer.

Concernant la sécurité des vaccins, une histoire visuelle en douze actes (Figure 21) comprenant des photographies, des tracés de zones, des cartes choroplèthes, des tableaux, un bref texte et d'autres composants a été créée. Une courte introduction de deux actes,

9. La thérapie narrative est une forme de psychothérapie qui cherche à aider les patients à identifier leurs valeurs et les compétences qui leur sont associées. Il permet au patient de connaître sa capacité à vivre ces valeurs afin de pouvoir affronter efficacement les problèmes actuels et futurs.(Wikipédia)

10. Selon l'Organisation mondiale de la santé (OMS), la littératie en santé correspond aux « aptitudes cognitives et sociales qui déterminent la motivation et la capacité des individus à obtenir, comprendre et utiliser des informations d'une façon qui favorise et maintienne une bonne santé »

une section médiane de neuf actes et une partie finale d'un acte composent cette histoire. Le public cible était constitué de parents sceptiques, indécis, totalement opposés ou préoccupés par les avantages de la vaccination.

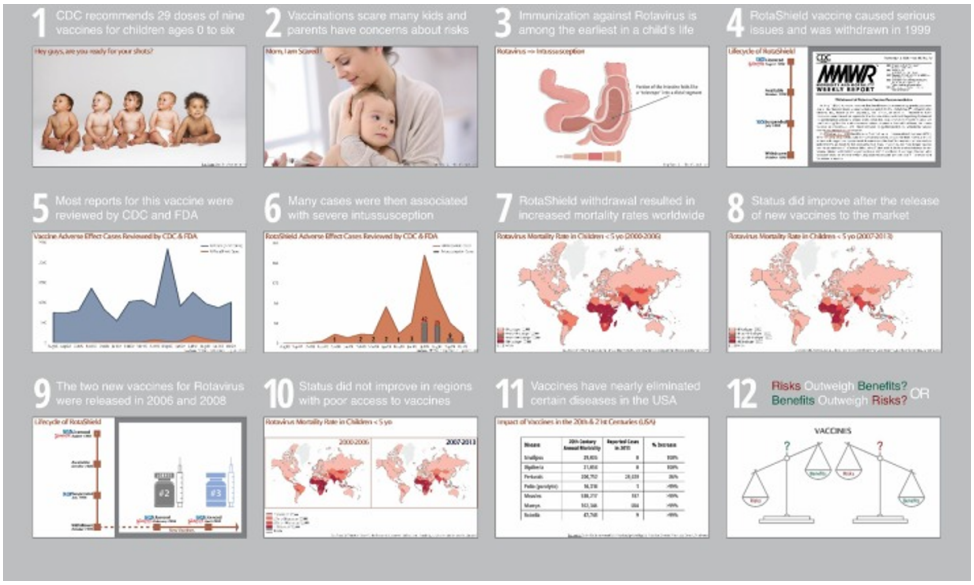


FIGURE 21 – Application du cadre théorique : actes histoire visuelle vaccin (Botsis et al., 2020)

Pour ce qui est de l'histoire sur l'immunothérapie du cancer, elle était structurée en 11 actes (une courte introduction de quatre actes et une section centrale de sept actes) qui incluent des illustrations médicales, des diagrammes de pente, des graphiques et du texte. Le public cible de cette histoire était les patients atteints d'un cancer du poumon qui ne connaissent pas l'immunothérapie, son mécanisme d'action et les défis liés à l'interprétation de la réponse à la thérapie basée sur l'imagerie.

Dans (Peddireddy et al., 2020), un tableau de bord (Figure 22) pour la surveillance de la COVID-19 a été développé. L'objectif de ce tableau était de fournir un outil visuel simple pour comparer, organiser et suivre les données de surveillance en temps quasi réel au fur et à mesure de la progression de la pandémie. À travers ce tableau de bord, il est possible de visualiser (a) une carte choroplèthe du monde, rendue avec l'estimation du nombre d'actifs, mais avec la possibilité de passer à d'autres couches ou à des attributs différents ; (b) un panneau d'informations qui comprend des graphiques interactifs, des statistiques sommaires et un tableau de données et (c) un panneau d'information qui présente l'analyse par réponse aux questions et comprend des résultats avec des données et un graphique.

Enfin, (Kleinau et al., 2022) ont produit une histoire de données pour informer le grand public sur l'utilisation des données de débit sanguin mesurées pour diagnostiquer et traiter les maladies cardiovasculaires. Cette histoire se concentre sur les tourbillons de flux sanguin dans l'aorte, avec quelle technique d'imagerie médicale, ils sont examinés, et

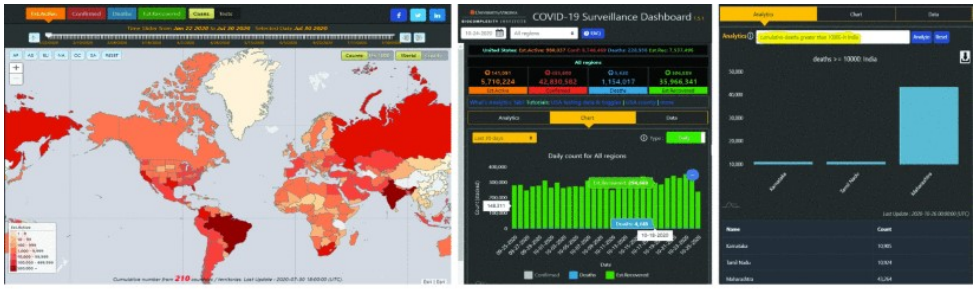


FIGURE 22 – Captures d'écran du tableau de bord COVID-19 (Peddireddy et al., 2020).

pourquoi ils peuvent être dangereux. La construction de l'histoire s'est fait de la manière suivante : (i) l'extraction et prétraitement des données, (ii) la définition des ingrédients de l'histoire (contenu et personnage de l'histoire), (iii) la définition des questions pour la conception de l'histoire (par exemple : combien de personnages doivent figurer dans l'histoire et quel est leur design (personnes réalistes ou dessins illustratifs) ?, Comment pouvons-nous transmettre le phénomène des tourbillons de l'écoulement sanguin ?) et (iv) le choix de la structuration de l'histoire. L'histoire est décomposée en 5 actes qui sont organisés selon la structure Martini Glass (Segel and Heer, 2010).

Comme leçon apprise dans la production de cette histoire visuelle de données, les auteurs affirment que l'expérience de communication des médecins avec leurs patients est particulièrement utile pour concevoir une visualisation médicale narrative.

2.3.1.4 Conclusion

Les travaux que nous avons pu étudier montrent qu'en dehors de (Peddireddy et al., 2020) qui a développé un tableau de bord pour la surveillance épidémiologique de la COVID-19 destiné à un public d'experts ainsi qu'aux citoyens, d'autres narrations produites sont exclusivement destinés au grand public pour la sensibilisation et l'éducation pour la santé. En outre, seule (Kleinau et al., 2022), donnent les phases suivies pour la construction d'une histoire sur l'utilisation des données de débit sanguin mesurée pour diagnostiquer et traiter les maladies cardiovasculaires.

Après l'étude des processus de narration de données, nous passons maintenant à l'étude des processus en intelligence épidémique.

2.3.2 Processus d'intelligence épidémique

Dans cette sous-section, après une présentation des concepts de base sur l'intelligence épidémique, nous réalisons un état de l'art sur les processus d'intelligence épidémique.

2.3.2.1 Définition des concepts

L'intelligence épidémique réfère à la détection précoce des épidémies (maladies transmissibles) ou à l'ensemble des problèmes de santé publique. En intelligence épidémique, le système de surveillance peut être basé sur les indicateurs ou sur les évènements.

Concernant la surveillance basée sur les indicateurs, l'Organisation Mondiale de la Santé (OMS) l'a défini comme : « le recueil systématique (régulier), le suivi, l'analyse et l'interprétation de données structurées d'indicateurs-produits par un nombre bien identifié de sources formelles constituées principalement de structures sanitaires » (WHO, 2014). Les données utilisées dans ce type de système de surveillance proviennent principalement des structures sanitaires (par exemple, hôpitaux et laboratoires), mais aussi d'autres sources non médicales (par exemple, les données météorologiques). Aussi, la collecte des données suppose l'établissement de définitions de cas de maladies/syndromes, l'identification des sources d'information appropriées et le choix des fréquences et des mécanismes de transmission des données.

Dans une surveillance basée sur les indicateurs, les objectifs sont : (i) l'alerte précoce pour les pathologies sévères ou à potentiel épidémique (par exemple le choléra, Ebola, la grippe saisonnière) et (ii) le suivi à plus long terme (morbidité, suivi des programmes, planification) pour des pathologies infectieuses (par exemple la tuberculose), les maladies non transmissibles ou chroniques (par exemple le cancer, le diabète) et les expositions à des toxiques.

Pour ce qui est de la surveillance basée sur les évènements, elle est définie par l'OMS comme « le recueil organisé, le suivi, l'évaluation et l'interprétation d'information ad hoc, principalement non structurées, et concernant des évènements ou des risques, qui peuvent représenter une menace aiguë pour la santé humaine » (WHO, 2014). Les informations recueillies à travers ce système de surveillance sont de nature diverse et peuvent provenir des multiples sources, officielles ou non-officielles et souvent non préétablies (par exemple, les rumeurs provenant des médias). Ce système de surveillance doit œuvrer en complément de la surveillance basée sur les indicateurs qui est une surveillance de routine.

Dans le cadre de ce travail, le système de surveillance épidémiologique objet de notre étude est basé sur les indicateurs. Car, ce système repose sur une organisation composée de centres de diagnostic et de traitement qui assurent la collecte continue (en routine) des données lors de la prise en charge des patients tuberculeux avant de les transmettre aux bases d'épidémiologie et de lutte contre les endémies (BELE). Ces BELE assurent à leur tour la compilation de ces données et leur analyse avant de les transmettre à l'institut d'épidémiologie de lutte contre les endémies (IELE) et au Programme National de Lutte contre la tuberculose (PNLT).

Après avoir présenté les principaux concepts sur l'intelligence épidémique, nous allons maintenant passer en revue les processus d'intelligence épidémique proposés dans la

littérature.

2.3.2.2 Processus d'intelligence épidémique

De nombreux processus d'intelligence épidémique (IE) ont été proposés dans la littérature (Figure 23). Il y a d'abord celui recommandé par l'Organisation Mondiale de la Santé (OMS) pour la surveillance des maladies qui est organisé en cinq grandes phases (WHO, 2014) :

- Détection : cette phase comprend (i) la définition des modalités de collecte des données et des informations et (ii) l'exécution de la collecte des données. Dans une surveillance basée sur les indicateurs, la phase de détection consistera à définir le type et la modalité (par exemple le format de la collecte ou le mode et la fréquence de la transmission) des données de surveillance à recueillir.
- Triage : à cette phase, deux activités sont réalisées, (i) l'analyse et (ii) l'interprétation des données. Concernant l'analyse des données, elle consiste à en vérifier la qualité et à effectuer l'épidémiologie descriptive et analytique, c'est-à-dire à organiser les données en temps, lieu, personnes, et de les organiser selon les facteurs de risque. Pour visualiser ces données, des tableaux, graphiques et cartes sont souvent utilisés. L'interprétation des données, quant à elle, consiste à convertir les statistiques en informations pratiques et utiles.
- Vérification : cette phase consiste à vérifier et à compléter les informations disponibles auprès de sources fiables telles que les instituts de santé publique, les ministères (santé, agriculture, environnement, ...), les laboratoires des pays concernés, l'OMS, les réseaux régionaux, les ONG, les ambassades, la source à l'origine de l'information, les autorités locales, etc.
- Évaluation du risque : cette phase consiste à caractériser l'évènement et d'estimer l'importance du risque qu'il représente en matière de santé publique.
- Communication : cette phase consiste à communiquer les alertes détectées au travers de supports adaptés aux différents acteurs et autorités concernés : ministère de la santé et autres ministères (notamment en cas d'évènement d'origine non infectieuse), autorités sanitaires internationales (par exemple, OMS), réseau national de santé publique, laboratoires, structures de soins et le cas échéant le grand public.

En comparant le processus d'intelligence épidémique de l'OMS au processus narration de données classique, nous relevons que les quatre premières phases (détection, triage (analyse et interprétation des données), vérification et évaluation du risque) du processus de l'OMS correspondent à la phase d'exploration de données du processus de narration de données classique. Tandis que la phase communication correspond à la phase présentation.

Ensuite, d'autres processus d'intelligence épidémique ont été proposés. Nous décrivons ces processus en les comparant à celui recommandé par l'OMS.

Dans (Che and Desenclos, 2002; Noah, 2006; Kaiser et al., 2006; Astagneau and An-

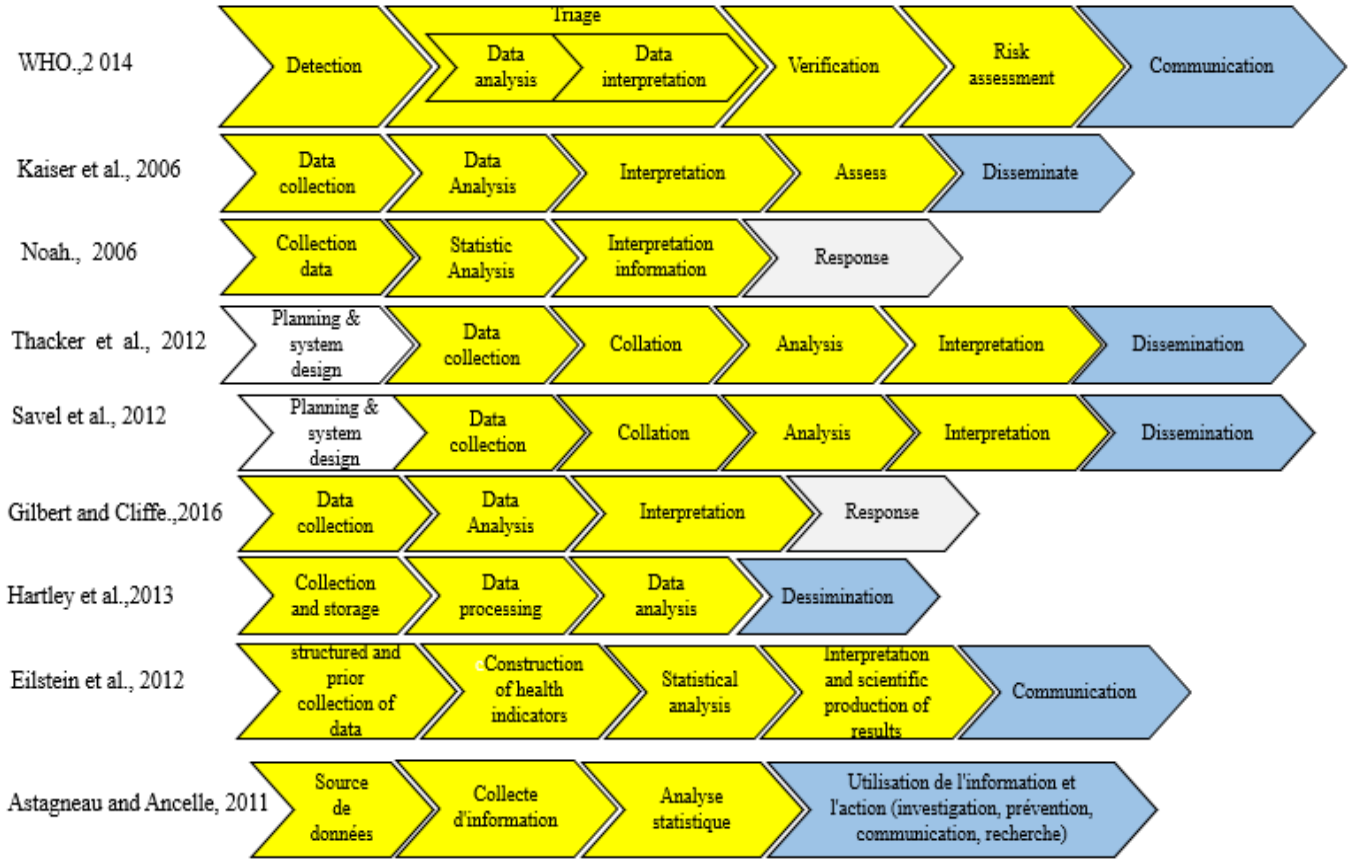


FIGURE 23 – Les processus d'intelligence épidémique proposés dans la littérature (WHO, 2014; Kaiser et al., 2006; Noah, 2006; Thacker et al., 2012; Savel et al., 2012; Gilbert and Cliffe, 2016; Hartley et al., 2013; Eilstein et al., 2012; Astagneau and Ancelle, 2011)

celle, 2011; Thacker et al., 2012; Savel et al., 2012; Eilstein et al., 2012; Hartley et al., 2013; Gilbert and Cliffe, 2016; Lourenço et al., 2019), bien que certaines phases des processus proposés soient nommées différemment (par exemple dissémination ou lieu de communication), les processus intègrent la majorité des phases ou des étapes (collecte des données, analyse des données, interprétation des données et communication) du processus de l'OMS. (Kaiser et al., 2006) ajoute la phase évaluation du risque épidémique et (Noah, 2006; Gilbert and Cliffe, 2016) une phase réponse (mise en place de politique de santé publique appropriées pour la riposte). Par contre, aucun des processus proposés ne propose la phase *vérification* du processus de l'OMS. Or, cette phase est très importante dans le processus d'intelligence épidémique.

Enfin, (Thacker et al., 2012; Savel et al., 2012) proposent une phase préalable, la planification et la conception du système de surveillance. Cette phase consiste à identifier les informations qui répondent le mieux à un objectif de surveillance et à déterminer qui aura accès à l'information (par exemple : décideurs ou grand public).

Dans cette section, nous avons présenté les concepts de base sur l'intelligence épidémique. Ensuite, nous avons réalisé une revue de la littérature sur le processus d'intelligence

épidémique. Cette étude a montré qu'en dehors de (Thacker et al., 2012; Savel et al., 2012) qui propose une première phase (planification et conception du système de surveillance) et (Noah, 2006; Gilbert and Cliffe, 2016) une phase mise en place de politiques sanitaires pour la riposte, tous les processus proposés s'alignent au processus d'intelligence épidémique recommandé par l'OMS.

2.3.3 Conclusion

De façon générale, dans le domaine de la santé publique, nous avons relevé qu'il existe peu de travaux sur la narration de données qui visent à transmettre des résultats scientifiques à un public expert afin de rendre compte de la situation sanitaire et des résultats des politiques de santé mises en œuvre et, plus généralement, d'aider à la prise de décision. De plus, les narrations de données produites, bien que souvent destinés au grand public, sont construites selon un processus ad hoc. Il y a un manque de processus pour la narration des données en santé publique. Dans ce travail, nous proposons un processus spécifique de narration des données pour l'intelligence épidémique. Ce processus personnalise les processus généraux de narration des données en intégrant les caractéristiques spécifiques de l'épidémiologie, les meilleures pratiques en matière d'intelligence épidémique et la communication aux épidémiologistes et aux autorités sanitaires.

2.4 Systèmes de simulation multi-agents

Afin de mieux modéliser et comprendre des épidémies (COVID-19, grippe, tuberculose, etc), il est fréquent de réaliser des simulations. Concernant la tuberculose, la grande majorité des modélisations proposées sont des systèmes classiques (SIR, SIS, SEIR) couramment utilisés en épidémiologie mathématique. De plus, ces modèles simulent seulement la dynamique de propagation de la maladie avec des pas temps de simulation relativement courts. Or, pour envisager circonscrire cette maladie dans une zone géographique donnée, au-delà de la compréhension de la dynamique de propagation, il faut également chercher à évaluer les politiques sanitaires de lutte.

C'est pourquoi dans ce travail, pour aider les experts de la santé publique qui sont des décideurs à évaluer les politiques sanitaires de lutte antituberculeuse (par exemple : la réduction du taux de perdus de vue, l'augmentation du nombre d'individus en traitement, etc.), nous cherchons à construire un modèle de simulation à base d'agents de la tuberculose réaliste. La construction d'un tel modèle nécessite que soit pris en compte la quasi totalité des états d'un individus dans le cycle de transmission et de prise en charge de cette maladie. Aussi, pour bien mesurer l'impact que pourrait avoir les différentes politiques sanitaires mises en oeuvre, les simulations doivent être réalisées à une échelle spatiale plus large (plusieurs communes ou arrondissements), sur des durées plus longues

(de plusieurs années) et avec des données réelles des personnes diagnostiquées positives à la tuberculose.

Cette section présente une synthèse de l'état de l'art sur les systèmes de simulation. Elle est organisée en deux sous-sections. Dans la première, nous définissons les concepts et présentons les différentes approches de modélisation. Pour chacune des approches, nous donnons un exemple d'utilisation en épidémiologie. A la deuxième, nous décrivons le système de simulation multi-agents et présentons les différentes plate-formes de simulation à base d'agents. Nous terminons par une conclusion.

2.4.1 Généralités sur la modélisation et la simulation

2.4.1.1 Modélisation au niveau macro et micro

La modélisation est un processus par lequel on organise les connaissances portant sur un système donné (Zeigler et al., 2000). Tandis qu'un modèle est une image simplifiée de la réalité qui sert à comprendre le fonctionnement de ce système en fonction d'une question (Bousquet et al., 2002). Autrement dit, il s'agit de la construction d'une représentation simplifiée et observable du comportement et/ou de la structure d'un système réel afin de résoudre un problème d'analyse ou de conception. Il existe plusieurs types de modélisations parmi lesquels, les modélisations prédictives, au niveau macro et au niveau micro.

La modélisation prédictive :

Les modèles prédictifs sont ceux qui permettent de décrire l'évolution d'un phénomène de manière à pouvoir prédire un état futur à partir d'un état initial connu. Ils visent uniquement à faire des prévisions sans essayer d'expliquer un phénomène (Bommel, 2009). Il s'applique généralement à des phénomènes à évolution rapide dans le temps et l'espace, comme des phénomènes épidémiologiques telle la tuberculose, ébola, la Covid-19, la bilharziose,... etc. La conception d'un modèle prédictif peut se faire à un niveau macro ou micro.

La modélisation niveau macro :

L'approche de modélisation macro se focalise sur une population d'agents dont l'architecture est simple, ce qui permet à cette population d'agents de résoudre un problème comme une seule entité le ferait (Grouls, 2013). Elle est considérée comme étant le fruit de la dynamique issue des interactions qui se déroulent entre les entités du niveau microscopique (Michel, 2006).

La modélisation niveau micro :

Une approche de modélisation micro se caractérise par deux éléments : d'une part le recours à des données très détaillées, a priori individuelles, d'autre part la modélisation des comportements individuels (Boulanger and Bréchet, 2003).

Dans ce travail de thèse, nous souhaitons mettre à la disposition des autorités sanitaires un outil qui leur permettra de prédire la propagation de la tuberculose en simulant de politiques sanitaires de riposte antituberculeuse. De plus, étant donné que notre modèle est centré-individu et que nous allons réaliser les simulations avec les données réelles des patients tuberculeux, l'approche micro est donc plus adaptée à notre contexte. Nous allons maintenant étudier comment concevoir une telle simulation.

2.4.1.2 Conception d'une simulation

La simulation est l'un des outils d'aide à la décision les plus efficaces à la disposition des concepteurs et des gestionnaires des systèmes complexes. Elle consiste à construire un modèle d'un système réel et à conduire des expériences sur ce modèle afin de comprendre le comportement de ce système et d'en améliorer les performances. La conduite d'une étude de simulation comprend trois phases (Figure 26) (Grimaud, 1998)) : l'analyse du problème, la construction simulation et l'exploitation simulation.

La phase d'analyse du problème est décomposée en deux étapes :

— Identification du problème ; spécification des objectifs ;

— Réalisation d'une première ébauche du modèle qui a pour but d'en délimiter les frontières et de spécifier les données dont on a besoin ;

— Validation auprès de l'utilisateur (celui qui est à l'origine de l'étude).

Le but visé est de construire un modèle valide qui soit le plus simple possible, tout en restant cohérent avec les objectifs de l'étude.

La phase de construction de la simulation comprend la modélisation logico-mathématique qui peut être facilitée par un outil graphique, et la programmation proprement dite.

Enfin, à la phase d'exploitation de la simulation, quand le modèle est validé, il peut servir à l'évaluation du comportement dynamique du système. Cette phase nécessite une définition précise de la campagne d'exploitation (quelles hypothèses veut-on vérifier, dans quel contexte), la production de mesures par la simulation proprement dite, la mise en forme et la comparaison des résultats obtenus aux objectifs poursuivis. S'ils n'ont pas été atteints, de nouveaux scénarios sont proposés et testés jusqu'à satisfaction. Aussi, dans la mesure où la plupart des modèles comportent des aléas, cette étape nécessite que soient déterminés avec rigueur la durée de la simulation et le nombre de réplications (exécutions du modèle de simulation).

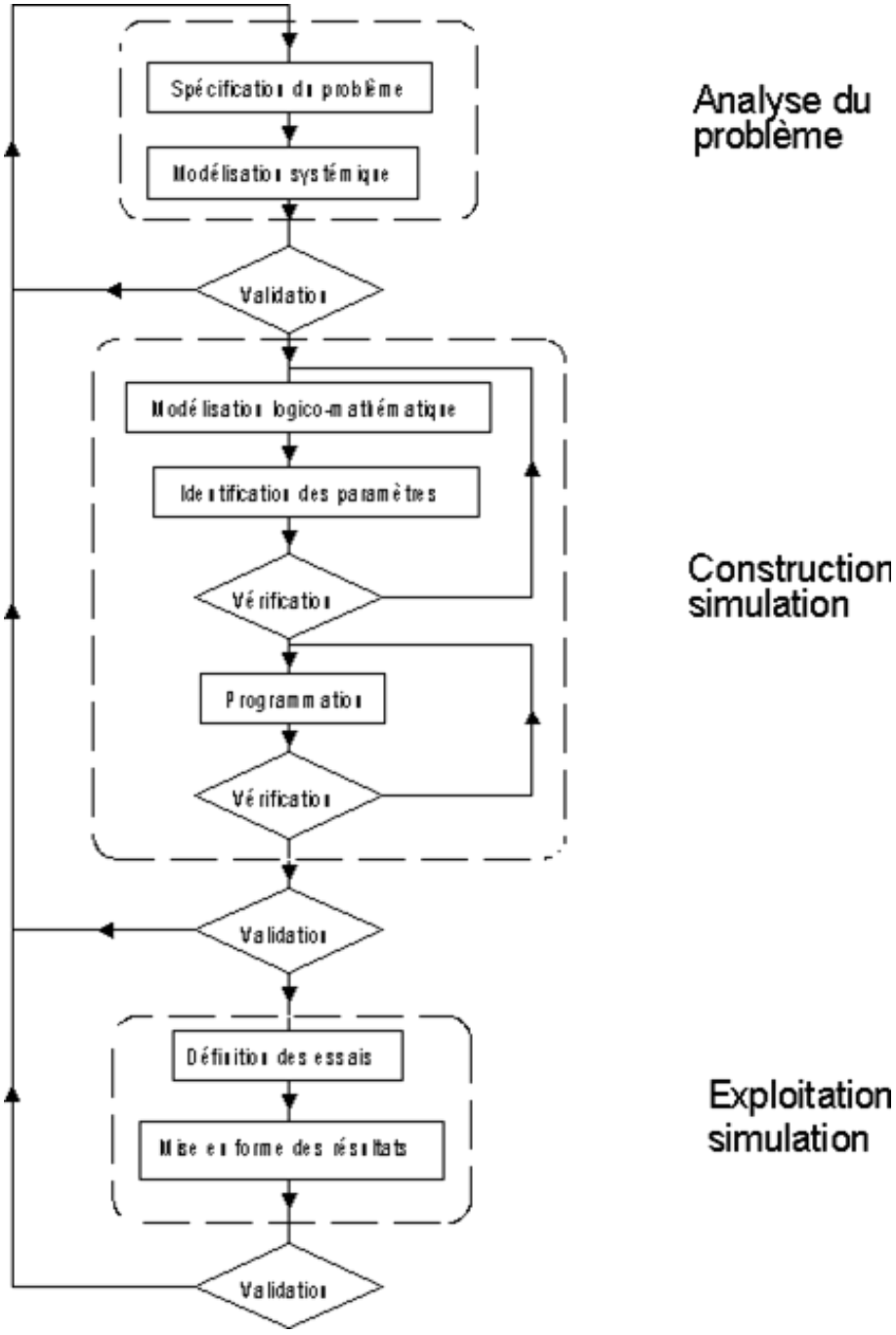


FIGURE 24 – Logigramme : étapes d'une étude de simulation (Grimaud, 1998)

2.4.1.3 Les différentes approches de modélisation

Il existe différentes approches de modélisation. Mais, les modélisateurs en épidémiologie ou en écologie généralement considèrent trois paradigmes de simulations : déterministe, stochastique et multi-agents (Bui et al., 2016). Ci-après, nous décrivons chacune de ces approches de modélisation et réalisons une analyse comparative de ces différentes approches.

Approche déterministe :

Le modèle déterministe est souvent le premier outil que va utiliser l'épidémiologiste pour étudier une nouvelle épidémie afin de déterminer la courbe globale d'infection et de trouver les meilleures valeurs des paramètres du modèle qui minimisent l'écart entre la courbe théorique et les valeurs observées (Bui et al., 2016). Cette approche de modélisation consiste à s'intéresser à l'évolution macroscopique des différentes populations, sans regarder à l'échelle de l'individu. Par exemple, le modèle SIR en épidémiologie déterministe décrit l'évolution dans le temps du nombre de personnes dans chaque compartiment et part du schéma simplifié suivant (Figure 25) (Villani, 2020) : un individu est d'abord sain (S) ; il est infecté (I) avec une probabilité beta, représentant le taux de transmission de la maladie d'une personne infectée à une personne saine ; enfin, il guérit (R), avec la probabilité gamma, représentant le taux de guérison, c'est-à-dire l'inverse de la durée moyenne des symptômes. Lors de l'épidémie de la COVID-19 qui a débuté en France le

FIGURE 25 – Modèle SIR (Villani, 2020)

24 janvier 2020, pour estimer le nombre réel de personnes infectées durant la période d'observation et d'en déduire le taux de mortalité associé à l'épidémie, le modèle SIR a été utilisé par (Roques et al., 2020). Les résultats ont montré que le nombre réel d'infectés en France était bien supérieur aux observations et que le taux de mortalité au 17 mars était de 5.2/1000 (IC [11] 95% 1.5/1000 11.7/1000).

Certaines épidémies se caractérisent par une période d'incubation. C'est une période durant laquelle un individu est infecté, sans présenter de symptôme ni être contagieux (Infectieux). Pour le prendre en compte, on ajoute le compartiment des exposés (E) au modèle SIR. On obtient ainsi le modèle SEIR (Brahim, 2019).

Approche stochastique :

Contrairement au modèle déterministe, un modèle stochastique est un modèle probabiliste permettant d'étudier un phénomène aléatoire au cours du temps. Il permet de décrire la réalité sous la forme d'un ensemble de paramètres numériques et d'un ensemble des relations mathématiques qui décrivent la manière dont certains de ces paramètres, appelés causes, agissent sur d'autres, appelés effets (Drogoul, 1993). Selon (Khuu, 2004), , en raison du problème de l'explosion du nombre d'états, elle est presque la seule approche

11. L'intervalle de Confiance(IC) à 95% est l'intervalle de valeur qui a 95% de chance de contenir la vraie valeur du paramètre estimé.

accessible permettant d'utiliser les formalismes de type états/transitions (par exemple une chaîne de Markov) pour la modélisation. Comme le montre la figure 26, dans un modèle stochastique, le phénomène réel se trouve modélisé sous la forme d'un certain nombre de variables-causes (a, b, c) et de variables-effets α, β, γ mises en relation par un ensemble de fonctions mathématiques F.

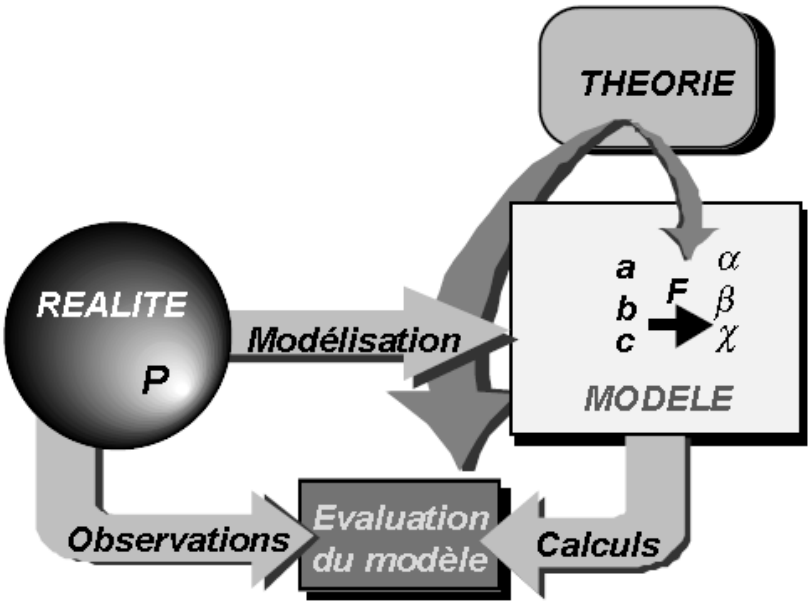


FIGURE 26 – Principes de la simulation stochastique (Drogoul, 1993)

Il existe plusieurs types de modèles stochastiques. Parmi ces modèles, nous pouvons citer :

— Les modèles compartimentaux : ils consistent à diviser la population en plusieurs compartiments, avant de décrire les interactions entre eux, par exemple à travers un système d'équations différentielles (Brahim, 2019).

— Les chaînes de Markov : une chaine de Markov est un processus stochastique où le nombre d'états (valeur d'une variable) est défini et l'état en t+1 dépend uniquement de l'état en t. Le passage d'un état à un autre est défini par une matrice de transition P (xi, xj) qui décrit la probabilité de passé de l'état xi en t à l'état xj en t+1.

La figure 27 présente un exemple de chaîne de Markov. Les flèches indiquent les probabilités de transition. la somme des probabilités des flèches partant de chaque état doit être égale à 1.

En somme, les modèles déterministes rendent compte d'un comportement moyen observé un niveau macroscopique. Ils décrivent notamment l'évolution des concentrations des différentes espèces, par des équations différentielles dont les champs de vecteurs sont les débits de tous les évènements impliqués dans la dynamique (Nankep et al., 2018). Par contre, ce type de modèle ne permet pas de fournir des micro-spécificités. Nous allons

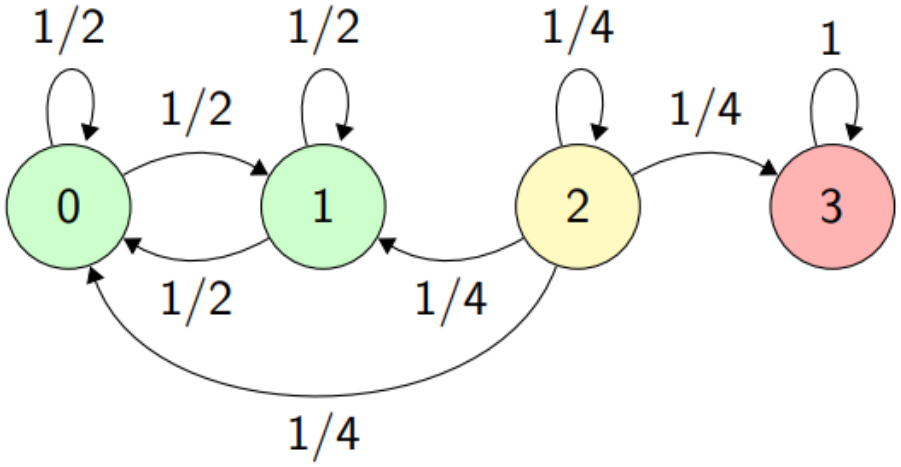


FIGURE 27 – Exemple d'une chaîne de Markov

maintenant décrire les modèles multi-agents qui eux autorisent la prise en compte de ces micro-spécificités.

Approche multi-agents :

L'approche multi-agents décompose un phénomène en plusieurs parties (les agents) et la dynamique globale du système résulte des interactions entre toutes ces parties et leur environnement (Siebert, 2011). La modélisation d'un phénomène dans une perspective multi-agents se fait en plusieurs étapes. Selon (Taillandier, 2019), ces étapes sont les suivantes (Figure 28) :

— *Définition de la question de modélisation* : un modèle doit être construit par rapport à une (ou plusieurs) question à laquelle on souhaite que le modèle puisse nous aider à répondre. Définir cette ou ces questions est ainsi une étape clef du processus de modélisation qui permettra de guider tout le reste du processus.

— *Identifier les éléments à modéliser* : cette seconde étape consiste à identifier par analyse du système réel (directement ou indirectement au travers d'experts ou de la littérature) les éléments (entités, dynamiques...) que l'on va devoir intégrer dans le modèle si l'on souhaite pouvoir répondre à la question de modélisation.

— *Collecter des données* : une étape importante, en particulier lorsqu'on développe un modèle particulier, et qui va souvent demander des ressources importantes, est la collecte des données. Ces données peuvent être de différentes natures aussi bien quantitatives que qualitatives et provenir de différentes sources (base de données géographiques ou statistiques, entretiens, dire d'expert...). Un travail de nettoyage, de croisement, voire de génération de données synthétiques sera souvent nécessaire à cette étape.

— *Définir les agents* : l'objectif de cette étape est d'aller plus loin dans la formalisation du modèle en concevant un modèle conceptuel complet. Ainsi, il s'agira ici de

définir quels seront les types d'agents utilisés dans le modèle, quelles seront leurs caractéristiques et leurs dynamiques.

— *Implémenter le modèle* : cette étape consiste à transformer le modèle conceptuel conçu dans l'étape précédente en un modèle implémenté. S'il est courant qu'il y ait un écart entre le modèle conceptuel et le modèle implémenté (ajout ou suppression de certains types d'agents, modification de la nature de certaines variables ou ajout d'autres variables pour diminuer les temps de calcul), il est important de toujours documenter les choix qui ont été faits afin que même ceux qui n'ont pas participé à cette phase d'implémentation puissent toujours comprendre ce qu'il y a dans le modèle.

— *Calibrer le modèle* : cette étape optionnelle consiste à définir les valeurs de paramètres, en particulier des paramètres sur lesquels peu de données sont disponibles, permettant au modèle de reproduire au mieux le système réel. Elle nécessite de disposer de données ou de connaissances sur le système étudié. Si le modélisateur dispose d'une métrique pour caractériser numériquement la validité d'une simulation (par exemple, écart entre les données observées sur le système réel et les données simulées), il est possible de recourir à des méthodes de calibration automatique qui vont explorer les combinaisons possibles de valeurs de paramètres afin de trouver celles optimisant au mieux la métrique. On cherchera aussi dans cette étape à vérifier l'impact des différents paramètres du modèle sur les sorties de simulation.

— *Explorer le modèle* : cette étape consiste à utiliser le modèle pour répondre à la question de modélisation.

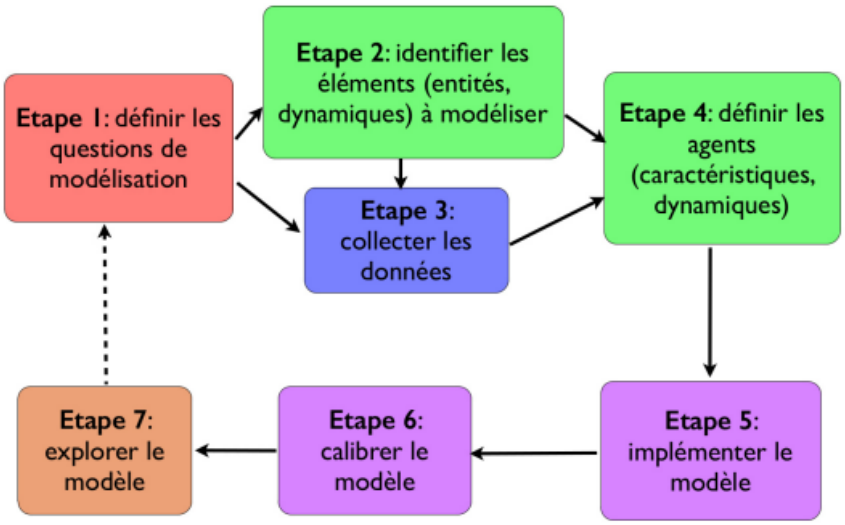


FIGURE 28 – Cycle de modélisation (Taillandier, 2019)

En ce qui concerne les maladies infectieuses, les systèmes de simulation à base d'agents

permettent de : (i) individualiser les populations (en représentant chaque individu sous forme d'agents) afin de prendre en compte leurs actions individuelles et les interactions entre elles qui sont primordiales dans la transmission d'une maladie infectieuse d'une personne à une autre et (ii) intégrer les dimensions spatiale et sociale spécifiques à la propagation des épidémies (Cisse, 2016).

Dans la littérature, de nombreux travaux (Badariotti et al., 2007; Sharma et al., 2020; Ziyadi et al., 2008; Kamla, 2008; Cisse et al., 2017; Prats et al., 2016; Vila Guilera, 2017; Balama and Corneille, 2017; Kabunga, 2020) utilisent l'approche à base d'agents pour modéliser la propagation des maladies infectieuses. Nous faisons un focus sur les travaux qui modélisent et simulent la propagation inter-individus de la tuberculose en utilisant l'approche multi-agents.

Parmi ces travaux, nous pouvons d'abord citer (Prats et al., 2016). Dans ce travail, un modèle de système multi-agents pour analyser l'évolution de l'incidence de la tuberculose pulmonaire au sein d'une communauté d'un quartier (Ciutat Vella) de la ville de Barcelone en Espagne est proposé. Ce modèle (Figure 29) est composé d'individus sains, infectés (tuberculose latente), malade (tuberculose active), en traitement et susceptibles d'être malade, infecté ou sain. Dans ce modèle, il ne ressort pas clairement les différents états (issu thérapeutique) que peut prendre un individu en traitement. En effet, dans le cadre de la prise en charge de la TB, un individu mis sous antituberculeux peut avoir *achevé son traitement* sans forcément avoir été déclaré guéri dans son dossier médical. Aussi, les états, *interruption de traitement*, *échecs de traitement*, *perdu de vue* sont des états que peut prendre un individu mis sous traitement antituberculeux. Mais, les auteurs ne les ont pas considérés dans leur modèle.

Pour améliorer le suivi de l'épidémie de tuberculose par les autorités sanitaires, le modèle qui doit être utilisé doit intégrer l'ensemble de ces états.

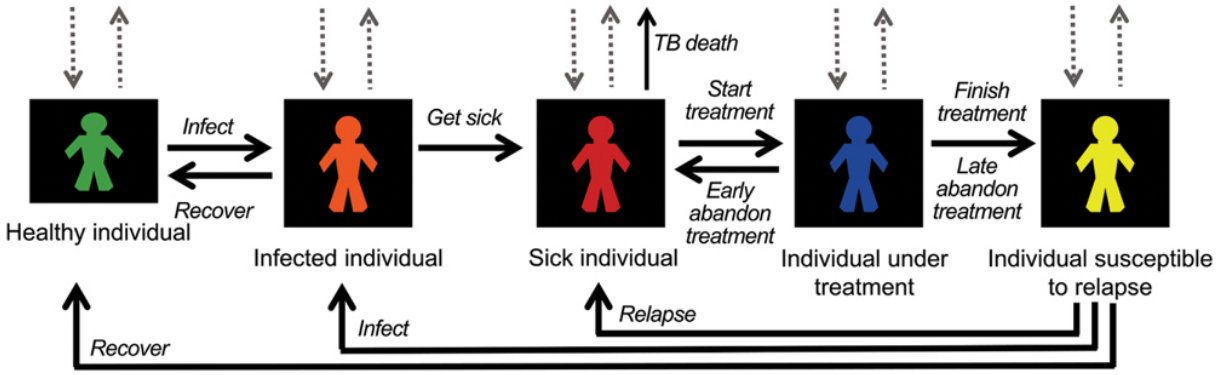


FIGURE 29 – Diagramme d'état du modèle (Prats et al., 2016)

Le modèle de (Prats et al., 2016) a été ensuite enrichi par (Vila Guilera, 2017) en prenant en compte les paramètres socio-démographiques (âge, sexe, origine) de la population. Les auteurs ont implémenté leurs modèles sous la plate-forme de simulation multi-agents

NetLogo.

Enfin, dans (Balama and Corneille, 2017), les auteurs proposent un modèle basé sur un SMA de la propagation de la tuberculose dans la ville de Ngaoundéré au Cameroun. Ils considèrent que la contamination se fait lorsque les individus susceptibles d'être infectés se déplacent dans un environnement (du lieu de résidence vers l'école, le marché, l'église, la mosquée, le travail et l'hôpital) qui contient une certaine quantité de Bacille de Koch dans l'air. Dans ce modèle, un individu peut seulement transiter par trois états, susceptible, infecté et malade. Donc, comme le modèle proposé par (Prats et al., 2016), celui-ci présente également quelques limites. Car, il ne prend pas en compte les états liés au résultat thérapeutique ou issu du traitement (guérison, traitement achevé, traitement interrompu, perdu de vue) du patient malade. Or, il est important de prendre en compte ces différents états. Car, ils permettent de mieux comprendre la dynamique de propagation la maladie (par exemple : les perdus de vue qui ensemencent leur entourage dissémine la TB) et de sa prise en charge. Les auteurs ont utilisé le système d'information géographique (SIG) de la ville de Ngaoundéré (fichiers au format Shapefile) comme environnement de simulation. Le modèle a été implémenté et simulé avec la plateforme GAMA.

2.4.1.4 Comparaison des différents niveaux et approches

La modélisation au niveau macro est le plus souvent liée à l'approche déterministe avec des équations. Tandis que la modélisation au niveau micro est plus souvent liée à l'approche stochastique ou à base d'agents. Le tableau 2.4 extrait des travaux de (Abrami et al., 2014) présente la comparaison de deux approches de modélisation de systèmes multi-agents : l'approche macro et l'approche micro. Pour comparer ces deux approches de modélisation, les auteurs ont retenu quatre critères essentiels : (i) l'échelle de description (population ou individu), (ii) le type de modèle (mathématique ou multi-agents), (iii) le type d'environnement et (iv) le type d'interaction.

Table 2.4 – Comparaison approches de modélisation macro et micro (Abrami et al., 2014)

Approche macro	Approche micro
Densité de la population	Individus
Équation	Règles définissant les comportements (Si-alors, boucle, actions séquencées)
Aspatial	Spatialement explicite
Homogénéité des variables d'état	Hétérogénéité des individus (âge, couleur, taille, position spatiale). Interaction locale entre les individus et avec l'espace

Au regard des différents critères de chacun des modèles, le modèle micro ou distribué,

c'est-à-dire à base d'agents, répond mieux à notre besoin de modélisation de la transmission et de la prise en charge de la tuberculose. Car, ce modèle est individu-centré, c'est-à-dire qu'il intègre les règles de comportement à l'échelle de chaque individu (agents) et des interactions entre ceux-ci. Aussi, l'environnement est spatialement explicite, ce qui est important pour un système de simulation en épidémiologie comme celui que nous proposons. Enfin, les systèmes de simulation à base d'agents présentent l'avantage de pouvoir modéliser des situations complexes dont les structures globales émergent des interactions entre individus, c'est-à-dire de faire surgir des structures du niveau macro à partir de modélisations du niveau micro, brisant ainsi la barrière des niveaux si criante dans les modélisations classiques (Ferber, 1997).

La sous-section 2.4.2, présente les concepts des systèmes multi-agents et les plateformes de simulation à base d'agents.

2.4.2 Composants et plateformes des systèmes multi-agents

2.4.2.1 Différents composants d'un système multi-agents

Dans la littérature, nous avons retrouvé diverses descriptions des composants des systèmes multi-agents. À titre d'exemple, (Ferber, 1995) décrit un SMA comme un système composé de (Figure 30) :

- Un environnement (E) ;
- Un ensemble d'agents (A) qui représentent les entités actives du système proposé ;
- Un ensemble d'objets (O) auxquels on peut associer une position dans E à un moment donné. Ces objets peuvent être créés, détruits et modifiés par des agents ;
- Un ensemble de relations (R) qui unissent les objets et agents entre eux.

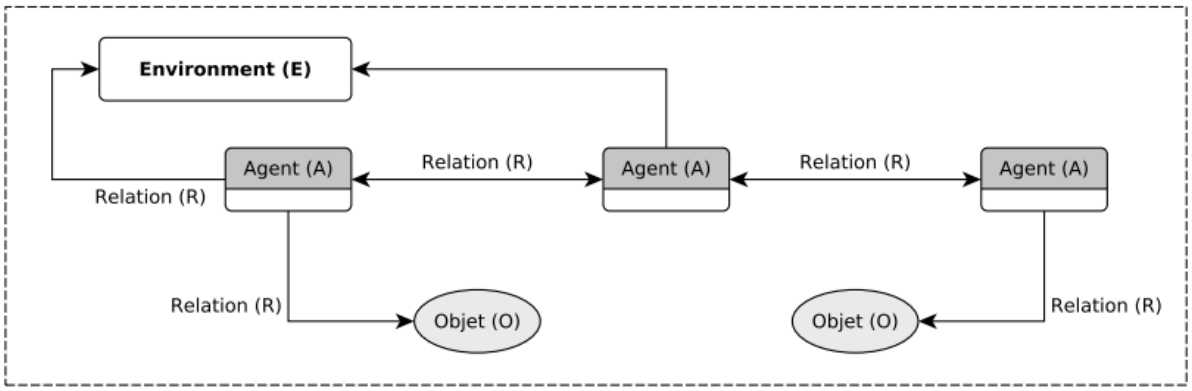


FIGURE 30 – Structure d'un SMA (Ferber, 1995)

Par contre, (Jennings et al., 1998) considère un SMA comme « un programme informatique regroupant des petits programmes disposant d'une certaine autonomie, constituant des entités artificielles qui évoluent, communiquent et agissent dans un environnement,

qui n'est pas forcément spatial ».

Pour tenir compte du contexte de notre travail de thèse, nous adoptons la définition de (Ferber, 1995) qui, en plus de la notion d'agent, intègre celle d'environnement qui est essentielle dans la modélisation des maladies infectieuses.

Définition d'un agent :

Selon (Ferber, 1995), un agent est une « entité capable d'agir sur elle-même et sur son environnement, qui réagit à ses transformations et qui possède une représentation partielle de cet environnement ». Il peut également être considéré comme « une entité dynamique définit par un ensemble de perceptions (entrées) traitées par sa structure interne et produisant, grâce à des mécanismes qui caractérisent sa dynamique, un ensemble d'actions » (Hermellin, 2016). Les agents peuvent être classés en deux familles (Kabachi, 1999) :

— Agents cognitifs : les agents cognitifs disposent d'une base de connaissances comprenant les diverses informations liées à leurs domaines d'expertise et à la gestion des interactions avec les autres agents et leur environnement. Les agents sont généralement « intentionnels », c'est-à-dire qu'ils possèdent des buts et des plans explicites leur permettant d'accomplir leurs buts.

— Agents réactifs : contrairement aux agents cognitifs, les agents réactifs ne sont pas « intelligents » pris individuellement. Ils ne peuvent que réagir à des stimuli simples provenant de leur environnement, et leur comportement est alors simplement dicté par leur entourage sans que ces agents ne disposent d'une représentation des autres agents ou de leur environnement.

Il existe également une famille d'agents qualifiée d'hybride. Il s'agit des systèmes hétérogènes comportant des agents ayant deux types de comportements (cognitif et réactif) (Kabachi, 1999).

Dans le contexte de notre travail de thèse, le type d'agent qui correspond mieux aux agents utilisés dans le modèle SMA que nous proposons peut-être classé dans le type réactif. Car leur interaction n'est pas volontaire. De plus, dans le modèle actuel, il n'y a pas d'interaction. Car, la période de simulation est à un pas de six mois et l'environnement de simulation à l'échelle du quartier.

Concept d'environnement :

L'environnement est une représentation du monde dans lequel les agents se situent (Pesty et al., 1997). Dans la littérature, plusieurs types d'environnement SMA sont définis : social, spatial (ou physique), culturel, de communication, etc. Dans notre travail, nous nous intéressons à l'environnement spatial. Selon (Demange, 2012), dans un SMA, un environnement physique est un environnement spatial dans lequel des agents sont

immergés. Ces agents peuvent se déplacer, percevoir et agir dans cet environnement.

Dans le cadre de ce travail de thèse, notre environnement de simulation est la région sanitaire Libreville-Owendo-Akanda du Gabon. Il est composé de quartiers dans lesquels interagissent les individus de notre modèle.

2.4.2.2 Plateformes multi-agents

Le processus de choix d'une plateforme multi-agents nécessite la prise en compte d'un certain nombre de critères. Parmi ces critères, il y a le nombre d'agents que la plateforme permet de modéliser, les différents modèles d'environnement possibles, le degré d'interaction entre agents, la capacité d'intégrer et de traiter des données géographiques par couplage avec un SGBD spatial. Le critère open source est tout aussi important. Car, il donne la possibilité au développeur d'avoir accès au code puis de le modifier en cas de besoin pour l'extension du système. En nous appuyant sur ces critères, nous faisons un focus sur deux plateformes multi-agents : NetLogo et Gama. Ces deux plateformes remplissent la plupart des critères cités. Elles intègrent plusieurs modèles de démonstration (épidémiologie, biologie, etc.) et sont souvent mise à jour par la communauté. Ensuite, entre les deux plateformes, nous choisissons celle qui s'avère plus adaptée pour modéliser et simuler le système transmission et de la prise en charge de la tuberculose, en tant compte de la nécessité de réaliser des simulations à partir des données réelles stockées dans un entrepôt de données spatiales multidimensionnelles.

Netlogo :

Développé en 1999 par Uri Wilensky de l'Université de Northwestern (USA), la plateforme NetLogo utilise son propre langage de programmation (NetLogo) et son code source n'est pas accessible. Aujourd'hui, NetLogo permet d'importer des données géographiques, matricielles et vectorielles dans le format standard Shapefile des données. Cette fonctionnalité multiplie les possibilités de création de modèles d'agents géospatiaux. Cependant, NetLogo ne fournit pas des opérations d'analyses géographiques avancées (Othman, 2016).

NetLogo a été utilisée pour développer des applications dans plusieurs domaines, dont la biologie, les sciences sociales, l'épidémiologie, etc. Dans le domaine de l'épidémiologie, epiDEM (Epidemiology : Understanding Disease Dynamics and Emergence through Modeling) (Waggoner et al., 1969) est le modèle de base proposé. Il s'appuie sur le modèle en épidémiologie mathématique SIR (Susceptible-Infecté-Guéri) qui est connu sous le nom de modèle Kermack et McKendrick (Kermack and McKendrick, 1927). Dans ce modèle [12], on suppose une population fermée, c'est-à-dire qu'il n'y a pas de naissances, de décès ou de déplacements vers où hors de la population. Il suppose également qu'il y a un mélange homogène, c'est-à-dire que chaque personne dans le monde a la même chance d'interagir

12. Modèle epiDEM : https ://ccl.northwestern.edu/netlogo/models/epiDEMBasic

avec toute autre personne dans le monde. En ce qui concerne le virus, le modèle suppose qu'il n'y a pas de période de latence et ni de risque de mutation virale. Le modèle epiDEM est un modèle simple de type SIR qui facilite les analyses mathématiques et le calcul du seuil à partir duquel une épidémie est susceptible de se produire. La figure 31 présente la fenêtre de visualisation d'une simulation du modèle epiDEM avec NetLogo 6.0.4. L'interface du modèle epiDEM (Figure 31) est composée de paramètres d'entrée :

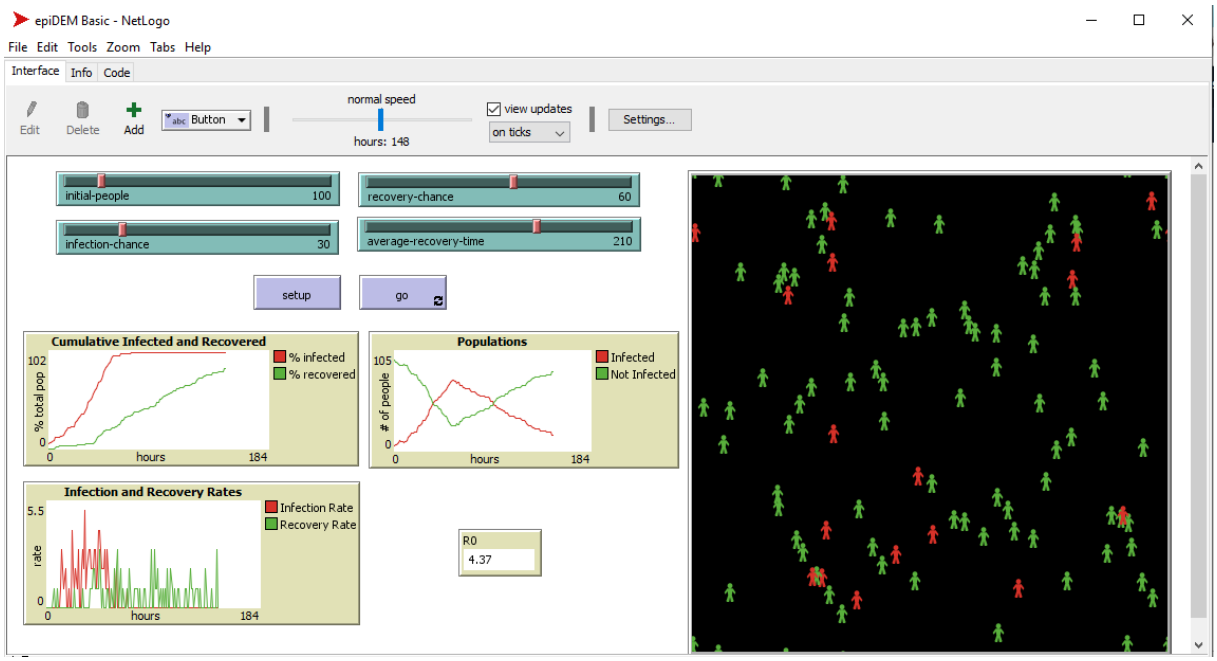


FIGURE 31 – Fenêtre de visualisation d'une simulation avec NetLogo : modèle basic epiDEM

— La population initiale (initial-people)
— Recovery-change : probabilité d'un individu d'être rétabli ;
— Infection-change : probabilité d'infection d'un individu ;
— Average-recovery-time : durée moyenne pour qu'un individu soit rétabli.

Les boutons *Setup* et *Go* permettent respectivement d'initialiser et de lancer la simulation. Après lancement de la simulation, l'utilisateur peut visualiser les graphiques de l'évolution du pourcentage d'individus infectés et rétablis (Cummulative Infected and Recovery), le taux d'infectiosité et de guérison (Infected and Recovery rate). Enfin, dans l'environnement, l'utilisateur peut visualiser les interactions entre individus infectés (en rouge) et rétablis (en vert).

En dehors des modèles exemples (epiDEM) proposés dans la plate-forme, NetLogo a été utilisé dans de nombreux travaux pour modéliser et simuler la propagation d'autres épidémies. Par exemple, dans (Kardas-sloma et al., 2019), elle a été utilisée pour modéliser et simuler la transmission d'Escherichia coli [13] résistant aux Béta-lactamines (E.

13. Escherichia coli, également appelée colibacille et abrégée en E. coli, est une bactérie intestinale des mammifères, en forme de bâtonnet, très commune chez l'être humain. (Wikipédia)

coli BLSE). Plus récemment, dans (Izewski, 2022), cette plateforme a été utilisée pour la conception d'un modèle de simulation COVID-19 pour déterminer de meilleures stratégies de gestion de l'épidémie.

La plate-forme NetLogo se limite à la construction de modèle simple ne pouvant prendre en compte qu'un nombre très faible d'agents. Par contre, GAMA permet la construction de modèles très complexes qui intègrent de nombreux agents et données (Taillandier et al., 2014).

Plate-forme GAMA :

Gama est une plate-forme open source de modélisation et de simulation à base d'agents développée depuis 2007 [14]. Elle permet de construire des simulations à temps discrets. Cela signifie que le temps est segmenté en un ensemble de pas de simulation ayant tous la même durée : à chaque pas de simulation, tous les agents sont activés et peuvent agir (modification de leur état, de l'état d'autres agents ou de l'état du monde) (Taillandier et al., 2014).

La plateforme Gama est également caractérisée par ses nombreux outils permettant l'intégration et la manipulation des données géographiques. Elle dispose de son propre langage de modélisation : GAML (langage de modélisation de GAMA). Il s'agit d'un langage orienté agent dédié à la définition de simulations basées sur des agents. Il prend ses racines dans les langages orientés objet comme Java ou Smalltalk. Avec GAML, la connexion à un entrepôt de données spatiales multidimensionnelles est faite directement dans le code en spécifiant les paramètres de connexion. Ainsi, l'accès à un entrepôt de données spatiales multidimensionnelles est considéré comme un agent capable d'exécuter des requêtes SQL.

Tout comme NetLogo, GAMA a été utilisée pour modéliser et simuler des phénomènes dans des domaines tels que l'épidémiologie, le transport, etc. Dans le domaine de l'épidémiologie, la version 1.8.1 propose des modèles compartimentaux SI et SIR. La figure 32 présente un exemple de résultat de simulation avec le modèle en épidémiologie SIR avec GAMA. L'interface de simulation est composée de (i) paramètres : population initiale, les probabilités de transition d'un individu de susceptible (S) à infecté (I) (S->I) et d'infecté (I) à rétabli (R) (I->R), (ii) l'environnement de simulation qui permet de visualiser les interactions entre agent S, I et R et (iii) les graphiques qui permettent de visualiser l'évolution des populations d'individus S, I et R.

D'autres modèles de simulation en épidémiologie ont été également construits à partir de la plateforme multi-agents GAMA. Nous pouvons citer le travail de (Balama and Corneille, 2017) qui propose un modèle de simulation de l'épidémie de tuberculose présenté à la sous-section 2.4.1.3 et celui de (Ban et al., 2020) qui permet de simuler l'épidémie de

14. https ://gama-platform.org/

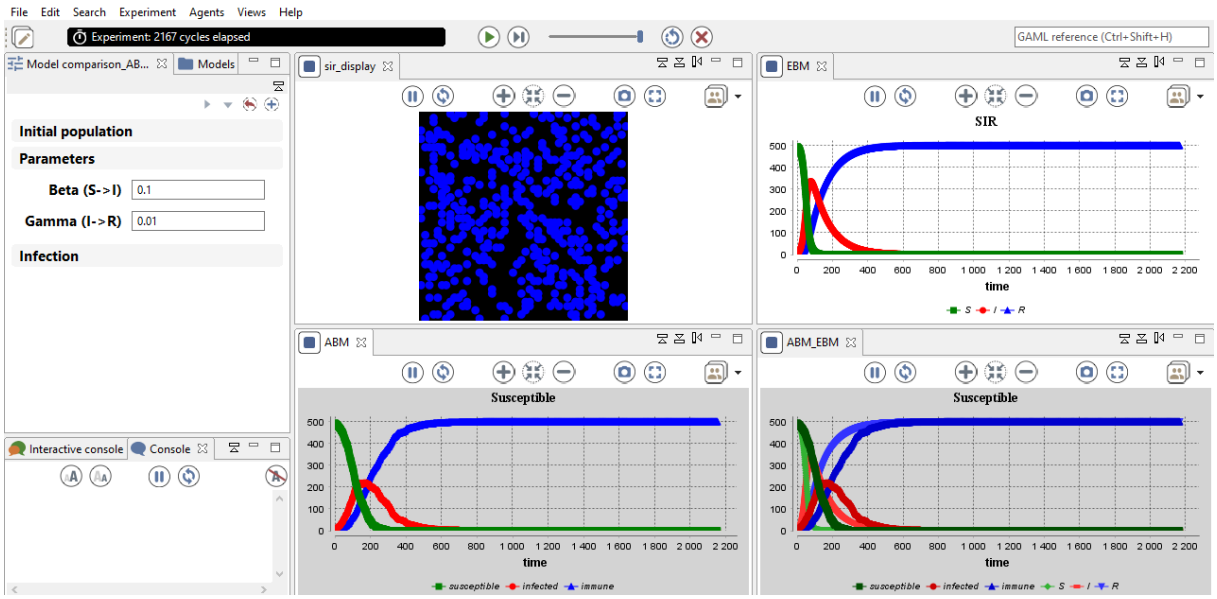


FIGURE 32 – Exemple résultat d'une simulation avec GAMA : modèle épidémiologie SIR

COVID-19.

Dans ce travail de thèse, pour la construction de notre simulateur, nous avons opté pour la plateforme GAMA. Notre choix a été guidé par deux critères que possède cette plateforme et qui permettent de couvrir notre besoin. En effet, GAMA permet la construction de modèles complexes qui font intervenir un grand nombre d'agents. Aussi, l'intégration des données (spatiales et descriptives) peut se faire par connexion à un SGBD spatial, ce qui reste encore difficile avec NetLogo. Or, dans le SMA que nous proposons, l'environnement géographique et les agents qui interagissent dans cet environnement doivent être générés et mise à jour à partir des données stockées dans l'entrepôt de données que nous concevons au niveau du SOLAP de la tuberculose (Section 3.1).

2.4.3 Conclusion

Dans cette section, nous avons présenté les différentes approches de simulation. L'approche multi-agents est plus pertinente pour la conception du modèle de simulation de la transmission et de la prise en charge de la tuberculose que nous proposons dans ce travail. Car, elle met l'individu au centre et tient compte de la dimension spatiale qui est très importante en épidémiologie.

Concernant la plate-forme de simulation, notre choix s'est porté sur la plate-forme de simulation multi-agents GAMA. Car, avec cette plate-forme, il est possible de réaliser des simulations avec des gros volumes de données descriptives et spatiales réelles par couplage avec un SGBD Spatial. Aussi, elle donne la possibilité de modéliser des systèmes complexes comportant un très grand nombre d'agents.

Enfin, il y a des propositions de systèmes de simulation à base d'agents de la tuber-

culose (Prats et al., 2016; Vila Guilera, 2017; Balama and Corneille, 2017), mais ils ne sont pas adaptés à l'évaluation de politiques sanitaires de riposte contre la tuberculose. Pour pouvoir évaluer les politiques sanitaires de ripostes, il faut détailler l'évolution de la maladie et du traitement en complétant les états déjà pris en compte dans la littérature (Susceptible, infectés, en traitement, guéris) par de nouveaux états (perdus de vue, malade, traitement interrompu, échec de traitement, traitement achevé). Il faut également donner la possibilité de changer le pas temporel (du jour à plusieurs mois) et l'emprise spatiale (passage d'un quartier à une grande agglomération). Enfin, les simulations réalisées doivent s'appuyer sur des données réelles des personnes infectées par la tuberculose. Ces améliorations sont indispensables pour l'évaluation des politiques de santé publique.

Dans ce travail, nous proposons un modèle de simulation à base d'agents de la tuberculose complexe. Ce modèle tient compte des états nécessaires pour évaluer les politiques de santé publique de lutte antituberculeuse. Aussi, par couplage avec le système d'information géographique décisionnel (SOLAP) de la tuberculose que nous proposons dans notre première contribution, les données utilisées pour les simulations de quelques politiques sanitaires de lutte antituberculeuse sont extraites de l'entrepôt de données spatiales multidimensionnelles. Par ailleurs, la durée du traitement antituberculeux étant de six mois, ces simulations se font à des pas de temps très long (le semestre). Enfin, elles couvrent un espace géographique très grand, la région sanitaire Libreville-Owendo-Akanda du Gabon.

Chapitre 3

Contributions

Dans le cadre de la surveillance de l'infection tuberculeuse, plusieurs acteurs interviennent (autorités sanitaires ou décideurs, experts en santé publique, épidémiologistes, etc.). Chacun de ces acteurs a des besoins d'analyse et de visualisation des données spécifiques.

D'abord, les experts de la santé publique ou les épidémiologistes ont besoin de réaliser des analyses exploratoires des données hétérogènes de la tuberculose stockées dans un entrepôt de données spatiales multidimensionnelles. L'analyse exploratoire des données est un mode d'analyse qui consiste à découvrir, explorer et détecter empiriquement des phénomènes dans les données (Jebb et al., 2017). Selon (Komorowski et al., 2016), l'objectif principal de ce type d'analyse est d'examiner la distribution, les valeurs aberrantes et les anomalies dans des données afin d'orienter les tests spécifiques à une hypothèse (par exemple : relations potentielles entre des variables). Cette analyse fournit des outils pour générer des hypothèses en visualisant et en comprenant les données, généralement par le biais d'une représentation graphique (Natrella, 2010). Aussi, ces acteurs souhaitent également simuler les politiques sanitaires de riposte antituberculeuse via un modèle prédictif. L'objectif visé par l'utilisation de ce modèle est d'anticiper sur le phénomène de propagation de la tuberculose.

Ensuite, pour mieux comprendre la dynamique de propagation de la tuberculose, les autorités sanitaires souhaitent suivre l'évolution dans l'espace et dans le temps des différents indicateurs de surveillance de la tuberculose via des tableaux de bord dynamiques (par exemple : navigation dans l'ensemble de données tout en conservant un niveau de lecture consolidé, clair et accessible.) et interactifs (par exemple : interactions entre graphiques, cartes et tableaux pour affiner les analyses).

Enfin, il y a un réel besoin d'utiliser un moyen efficace pour communiquer facilement à un public d'experts en santé publique qui sont des décideurs les découvertes ou trouvailles extraites des analyses de données sur la tuberculose.

Pour répondre aux besoins des différents intervenants de la lutte antituberculeuse, nous proposons comme pistes de solution, la conception et la mise en place d'un système d'information décisionnel qui couple un système d'information géographique décisionnel (SOLAP-TB) avec un système de simulation multi-agents (SMA-TB) pour la surveillance épidémiologique de la tuberculose. Le SOLAP-TB sera utilisé pour suivre et surveiller via des tableaux de bord dynamiques et interactifs l'évolution dans l'espace et dans le

temps d'indicateurs de la tuberculose. Tandis que le SMA-TB servira à simuler différentes politiques sanitaires de lutte antituberculeuse.

Nous proposons également un processus de narration de donnée en intelligence épidémique générique. C'est-à-dire réutilisable pour la production d'une narration de données sur n'importe quelle maladie. Enfin, nous produisons une narration de données sur la pandémie de tuberculose au Gabon qui s'appuie sur ce processus et sur le SOLAP-TB.

Ce chapitre présente nos trois contributions. À la section 3.1, le processus décisionnel de conception du système d'information géographique décisionnel de la tuberculose (SOLAP-TB) est décrit. La section 3.2 présente le processus de narration de données en intelligence épidémique proposé. Ensuite, une narration de données sur la tuberculose au Gabon qui s'appuie sur ce processus est produite. Nous terminons par la présentation du système de simulation de la propagation inter-individus de la tuberculose (SMA-TB) à la section 3.3.

3.1 Système d'information géographique décisionnel de la tuberculose, Syn@TB

3.1.1 Contexte

Le système de surveillance de la tuberculose du Gabon est très peu performant pour de nombreuses raisons. Parmi celles-ci, il y a le système de collecte, de traitement, d'analyse et de diffusion de données qui est essentiellement manuel avec l'utilisation par les différents acteurs (médecin, infirmier, décideur) impliqués dans la lutte antituberculeuse d'une multitude d'outils au format papier (par exemple : registre des cas de tuberculose, registre de laboratoire, dossiers médicaux des patients tuberculeux, rapport mensuel d'activités). Ce qui ne permet pas de garantir la bonne qualité de ces données hétérogènes et rend très difficile leurs intégration et exploitation pour la prise de décision.

Dans ce contexte, il apparaît donc qu'il existe un besoin urgent d'amélioration de la qualité des données hétérogènes de l'infection tuberculeuse ainsi que d'intégration de ces données stockées dans des outils papiers pour améliorer la prise de décision des autorités sanitaires dans la lutte antituberculeuse.

3.1.2 Objectif

Notre objectif est de proposer une solution numérique qui permet d'assurer la qualité des données hétérogènes de la tuberculose ainsi que l'intégration de ces données dans un entrepôt de données spatiales afin de faciliter la production de tableaux de bord dynamiques d'indicateurs spatio-temporels de l'infection tuberculeuse. Cet objectif représente

notre premier verrou scientifique baptisé système d'information géographique décisionnel de la tuberculose (SOLAP-TB).

Cette section décrit le processus décisionnel suivi pour la conception du système d'information géographique décisionnel de la tuberculose (SOLAP-TB).

3.1.3 Processus de conception et de réalisation du SOLAP de la tuberculose

La figure 33 présente les différentes étapes du processus de conception et de réalisation du SOLAP de la tuberculose, Syn@TB. Dans les sous-sections ci-après, pour chacune de ces étapes, nous présentons la méthodologie utilisée et les résultats obtenus.

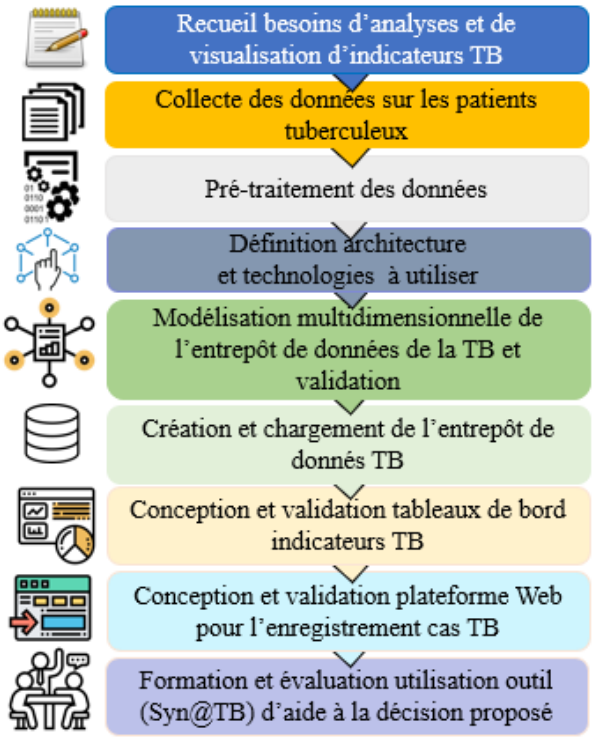


FIGURE 33 – Étapes de conception et de réalisation de Syn@TB

3.1.3.1 Recueil des besoins d'analyses et de visualisation d'indicateurs

La première étape de notre processus de conception a consisté à recueillir auprès des futurs utilisateurs (responsables du Programme National de Lutte contre la Tuberculose, directeurs régionaux de santé, chefs de bases d'épidémiologie et de lutte contre les endémies, responsables des centres de diagnostic et de traitement de la tuberculose) leurs besoins en termes d'analyses et de visualisation d'indicateurs.

Méthodologie : Nous avons réalisé une enquête. Un questionnaire a été conçu et validé. Une série de questions (cf. annexe B) ouvertes et fermées ont été posées à ces différents acteurs du système de surveillance de la tuberculose. Parmi les questions, il y avait celles qui concernaient (i) l'apport d'un système d'information décisionnel dans le renforcement du système de surveillance de la tuberculose, (ii) les outils utilisés pour la collecte des données, (iii) la qualité des données collectées dans l'actuel système de surveillance de la TB, (iv) l'utilisation faite des données reçues du système de surveillance de la TB et (v) les attentes en termes de renforcement du système de surveillance de la TB du Gabon.

Résultats : Au total, onze (11) personnes ont été enquêtées. Les réponses sont résumées au tableau 3.1. Les enquêtés ont tous répondu (100%) que la mise en place d'un système décisionnel pour la surveillance de la TB permettra une amélioration de la collecte, de la transmission et de la diffusion des données. Par contre, ils étaient très majoritaires (91%) à répondre que ce type de système permettra également d'améliorer l'analyse des données collectées lors de la prise en charge de patients tuberculeux.

Aussi, la très grande majorité (82%) des décideurs ont répondu que la qualité des données du système de surveillance de la TB est affectée par la tenue des supports de collecte de données (par exemple : registre des cas de tuberculose, dossiers médicaux au format papier) et une mauvaise collecte des données TB (par exemple : champs des dossiers médicaux des patients tuberculeux parfois mal renseignés ou non renseignés).

En termes d'attentes, tous les décideurs (100%) souhaiteraient que le système permette une analyse dans l'espace et dans le temps de la propagation de la TB. Tandis que 91% souhaitent consulter les données sous forme de tableaux de bord dynamiques et interactifs (cartes, tableaux et graphiques).

Conclusion : Les résultats de l'enquête ont montré que les acteurs impliqués dans la lutte antituberculeuse au Gabon ont un réel besoin d'un système d'information d'aide à la décision pour riposter efficacement contre cette maladie. Car, pour la très grande majorité d'entre eux, un tel système permettra d'améliorer la collecte des données, la qualité des données, l'analyse des données et la visualisation des données de surveillance.

3.1.3.2 Collecte des données

Méthodologie : Nous avons collecté les données sur les patients tuberculeux. Il s'agissait des patients qui ont été pris en charge entre 2016 et 2018 au centre de diagnostic et de traitement de la tuberculose de référence du Gabon : l'hôpital spécialisé Nkembo de Libreville (CDT-Nkembo). Une fiche de collecte de données a été conçue puis validée. Les dossiers médicaux des patients ont servi de principales sources de données. Les informations collectées concernaient les caractéristiques socio-démographiques, cliniques ainsi que le traitement. Le tableau 3.2 présente ces caractéristiques.

TABLE 3.1 – Résultats : état des lieux de l'existant et besoins des utilisateurs (décideurs)

Questions	**%**
Dans le contexte du Gabon, selon vous, que pourrait apporter la mise en place d'un système d'information décisionnel dans le renforcement du système de surveillance de la tuberculose ?	
Une amélioration de la collecte, la transmission et la diffusion des données ?	100%
Une amélioration de l'analyse des données ?	91%
Quels sont selon vous les éléments qui affectent la qualité des données du système de surveillance de la tuberculose ?	
La tenue des supports de collecte (registres) ?	82%
La mauvaise collecte de données ?	82%
Quelles sont vos attentes en termes de renforcement du système de surveillance de la tuberculose ?	
Analyse dans l'espace et dans le temps de la propagation de la tuberculose ?	100%
Consultation des données sous formes de cartes, tableaux et graphiques ?	91%

Ensuite, nous avons collecté les données géographiques de notre zone d'étude, la région sanitaire Libreville-Owendo-Akanda. Il s'agissait des fichiers formes (ou Shapefile en anglais) des quartiers, des six arrondissements et des trois communes.

Enfin, nous avons collecté les données de la population des arrondissements et des quartiers de notre zone d'étude. Les résultats du recensement général de la population et du logement (DGS, 2013) du Gabon ont servi comme première source de données. Ces résultats ne donnant pas le détail de la population par quartier, nous nous sommes appuyés sur une deuxième source de données, la population par quartier de l'enquête budget de consommation qui a été réalisée en 1992 à Libreville par la direction générale de la statistique du Ministère de la Planification (MPEAT and DGSEE, 1992).

Résultats : Au total, nous avons collecté les données socio-démographiques et cliniques dans 7 968 dossiers médicaux de patients tuberculeux du CDT-Nkembo. Ces données ont été ensuite saisies sous Excel et traitées avec l'ETL Talend avant d'être importées dans l'entrepôt de données spatiales multidimensionnelles de la tuberculose que nous avons préalablement conçu.

Concernant les données de population par quartier, nous avons pu obtenir les données de 84 quartiers sur 119 que compte la région sanitaire Libreville-Owendo-Akanda, soit 70,59%.

Enfin, pour les données géographiques, nous avons recueilli les shapefile (fichier forme géographique) des limites administratives des quartiers, arrondissements et communes de la région sanitaire Libreville-Owendo-Akanda.

TABLE 3.2 – Caractéristiques socio-démographiques et cliniques des patients

Caractéristiques	**Valeur**
Année	2016 ou 2017 ou 2018
Genre	masculin ou féminin
Statut professionnel	statut professionnel de chaque patient
Lieu de résidence	lieu de résidence de chaque patient
Type de tuberculose	pulmonaire, extra-pulmonaire et résistante aux antibiotiques (TBMR)
Type de patient	nouveau, rechute, traitement après échec, transfert entrant et autre cas traité (précédemment)
Traitement antituberculeux	RHZE (Rifampicine(R), Isosiazide(H), Pyrazinamide(Z), Ethambutol(E)) ou RHZ(Rifampicine(R), Isosiazide(H), Pyrazinamide(Z)), etc.
Résultat thérapeutique	guérison, traitement achevé, échec thérapeutique, décès, traitement interrompu, perdu de vue
Statut VIH	positif ou négatif
Date de début de traitement	date à laquelle le patient tuberculeux a commencé son traitement antituberculeux
Nationalité	nationalité du patient tuberculeux
Catégorie de patients	nouveau cas, rechute, échec, traitement après interruption
date de fin de traitement	date à laquelle le patient tuberculeux a achevé son traitement antituberculeux

La collecte des données pour la réalisation de notre entrepôt de données spatiales a été une tâche très fastidieuse. Car, les données de population de tous les quartiers de notre zone d'étude n'étaient pas disponibles. Aussi, pour les données des patients tuberculeux, il a fallu collecter ces données dans tous les dossiers médicaux papiers couvrant notre période d'étude (2016 à 2018).

3.1.3.3 Prétraitement des données

La construction d'un entrepôt de données nécessite avant tout d'évaluer la qualité des données qui y seront stockées et de rentre explicite et non ambigu le sens de ces données. Car, cette qualité peut impacter les indicateurs calculés et utilisés pour la prise de décision dans la lutte contre les épidémies. C'est pourquoi, nous avons d'abord procédé au prétraitement de l'ensemble des données (socio-démographiques, cliniques, géographiques et population) avant de les utiliser pour la production des tableaux d'indicateurs spatio-temporels au niveau SOLAP-TB et la simulation au niveau du SMA-TB.

Données de la population :

Méthodologie : Le traitement des données de la population a concerné celles des quartiers. Nous avons extrapolé ces données en appliquant le taux d'accroissement annuel de la population de Libreville qui est de 1,3%. Pour faire correspondre cette population à celle du recensement général de la population et du logement, nous avons appliqué ce taux à chaque population de quartier sur 20 ans. Par contre, l'échantillon de cette enquête n'était constitué que de 71% (n=84) de quartiers sur 119 qui composent la région sanitaire Libreville-Owendo-Akanda, parce qu'à cette période, certains nouveaux quartiers n'étaient pas encore intégrés dans cette région sanitaire. Par conséquent, il fallait donc compléter les données de population des 35 quartiers restant en parcourant à nouveau la littérature. Ces données n'étant pas disponibles, nous avons d'abord opté pour une répartition homogène de la population. Mais, cette méthode s'est avérée limitée, car elle ne tenait pas compte de la typologie des quartiers de chaque arrondissement. C'est pourquoi, nous avons choisi d'utiliser la procédure IPF (ajustement proportionnel itératif). Cette procédure a été proposée par Deming et Stephan en 1940. Elle permet d'ajuster un tableau de cellules de données de manière que leur somme corresponde aux totaux sélectionnés pour les colonnes et les lignes (dans le cas d'un tableau bidimensionnel) du tableau. Les cellules de données non ajustées peuvent être appelées les cellules "de base" et les totaux sélectionnés peuvent être appelés les totaux "marginaux".

Une procédure IPF commence par une matrice initiale. Lors de l'initialisation, les cellules des 35 quartiers sont vides dans le tableau de contingence. Ensuite, à partir d'un processus itératif, les valeurs des cellules sont mises à jour. Le processus est répété jusqu'à ce que l'une des conditions d'arrêt soit remplie (précision souhaitée, itérations maximales) (Zargayouna, 2021).

Résultats : La procédure IPF pour l'estimation de la population des 35 quartiers n'ayant pas de population connue s'est faite en plusieurs étapes : nous avons d'abord défini une matrice initiale (Tableau 3.3). Cette matrice est organisée comme suit, en ligne la population par arrondissement à ajuster et en colonne celle par typologie de quartier (précaire, mixte et moderne). Aussi, au cours de cette itération, nous avons calculé la population à affecter au type de quartier moderne du 1er arrondissement de Libreville : Total population quartier moderne - total population quartier moderne du 2ème arrondissement d'Akanda.

Ensuite, à la première itération (Tableau 3.4), pour chaque arrondissement, lorsqu'un type de quartier n'est pas concerné par le problème de population, nous affectons la valeur 0. Par contre, pour le(s) type(s) de quartiers concernés, nous avons affecté la population de l'arrondissement (en ligne).

Après cette première itération, nous avons obtenu la deuxième matrice de départ (Tableau 3.5).

TABLE 3.3 – Matrice départ 1 : Population arrondissement par typologie de quartier (précaire, mixte et moderne)

Arrondissement	**Précaire**	**Mixte**	**Moderne**	**Population**
1er Arrondissement Libreville	?	?	?	45 930
1er Arrondissement Owendo	?	?	?	35 717
1er Arrondissement Akanda	?	?	?	4 407
2ème Arrondissement Libreville	?	?	?	54 518
2ème Arrondissement Akanda	?	?	?	28 814
2ème Arrondissement Owendo	?	?	?	5 894
3ème Arrondissement Libreville	?	?	?	44 291
4ème Arrondissement Libreville	?	?	?	6 713
5ème Arrondissement Libreville	?	?	?	77 671
6ème Arrondissement Libreville	?	?	?	96 951
Population	266 235	95 645	39 021	400 906

TABLE 3.4 – Matrice itération 1

Arrondissement	**Précaire**	**Mixte**	**Moderne**	**Population**
1er Arrondissement Libreville	-	-	10 207	45 930
1er Arrondissement Owendo	-	-	0	35 717
1er Arrondissement Akanda	4 407	0	0	4 407
2ème Arrondissement Libreville	-	-	0	54 518
2ème Arrondissement Akanda	0	0	28 814	28 814
2ème Arrondissement Owendo	5 894	0	0	5 894
3ème Arrondissement Libreville	44 291	0	0	44 291
4ème Arrondissement Libreville	6 713	0	0	6 713
5ème Arrondissement Libreville	-	-	0	77 671
6ème Arrondissement Libreville	96 951	0	0	96 951
Population	266 235	95 645	39 021	400 906

Cette matrice a été complétée de la manière suivante, nous avons d'abord calculé le pourcentage de la population totale par type de quartier à compléter (précaire et mixte) en faisant le rapport entre la population totale par type de quartier (en colonne) sur la population totale à compléter (en colonne). Ensuite, les pourcentages trouvés ont été multipliés par la population par arrondissement en ligne, en tenant de la typologie des quartiers (Tableau 3.6).

À la troisième itération, nous avons réalisé un ajustement proportionnel par colonne

TABLE 3.5 – Matrice départ 2

Arrondissement	**Précaire**	**Mixte**	**Population**
1er Arrondissement Libreville	-	-	35 723
1er Arrondissement Owendo	-	-	35 717
2ème Arrondissement Libreville	-	-	54 518
5ème Arrondissement Libreville	-	-	77 671
Population	107 979	95 645	203 629

TABLE 3.6 – Matrice itération 2 : ajustement proportionnel en ligne

Arrondissement	**Précaire**	**Mixte**	**Population**
1er Arrondissement Libreville	17 862	17 862	35 723
1er Arrondissement Owendo	17 859	17 859	35 717
2ème Arrondissement Libreville	27 259	27 259	54 518
5ème Arrondissement Libreville	38 836	38 836	77 671
Population à compléter	107 979	95 645	203 629
total population temporaire	101 815	101 815	-

(Tableau 3.7). Cet ajustement s'est fait en faisant le calcul suivant : *Population ajustée en ligne x Population à compléter/ total population temporaire.*

TABLE 3.7 – Matrice itération 3 : Ajustement proportionnel en colonne

Arrondissement	**Précaire**	**Mixte**	**Population**
1er Arrondissement Libreville	18 943	16 779	35 723
1er Arrondissement Owendo	18 940	16 776	35 717
2ème Arrondissement Libreville	28 909	27 607	54 518
5ème Arrondissement Libreville	41 187	36 482	77 671
Population	107 979	95 645	203 629

Après convergence des données de la population, nous avons arrêté les itérations. Aussi, pour répartir dans chaque quartier les populations estimées à la première itération (Tableau 3.4) et la troisième itération (Tableau 3.7), nous avons divisé chaque population obtenue par le nombre de types de quartier concernés. Le tableau 3.8 présente le nombre de quartiers par type par arrondissement avec les données de la population non disponible.

Conclusion : Pour des besoins de l'étude, la procédure IPF nous a permis de compléter les cellules vides des données des populations des quartiers par ajustement proportionnel. Mais, un recensement complet récent de la population serait nécessaire pour avoir des données de population de meilleure qualité.

TABLE 3.8 – Nombre de quartiers par type par arrondissement avec les données de la population non disponible

Arrondissement	**Précaire**	**Mixte**	**Moderne**
1er Arrondissement Libreville	1	6	2
1er Arrondissement Owendo	9	2	0
1er Arrondissement Akanda	1	0	0
2ème Arrondissement Libreville	2	5	0
2ème Arrondissement Akanda	0	0	1
2ème Arrondissement Owendo	4	0	0
3ème Arrondissement Libreville	5	0	0
4ème Arrondissement Libreville	1	0	0
5ème Arrondissement Libreville	3	1	0
6ème Arrondissement Libreville	8	0	0

Données descriptives des patients :

3.1.3.3.1 Méthodologie : Pour analyser la qualité des données socio-démographiques des patients tuberculeux, nous avons utilisé le logiciel Talend Open studio 5.4.1. Les dimensions de la qualité des données évaluées étaient la complétude, l'unicité et la validité.

3.1.3.3.2 Résultats :

***Dimension complétude* :** Les résultats de l'évaluation de la complétude des données ont montré que les données concernant le lieu d'habitation étaient complétées à 100%. La complétude était quasi optimale pour les champs nom (99,90%), sexe (99,37%), âge (97,92%), nationalité (95,94%), date début du traitement (70,96%) catégorie du patient (98,31%), type de tuberculose (98,31%), type de patient (95,16%), résultat examen (61,18%) et prénom (59,68%). En revanche, la complétude était sous-optimale pour les champs de données date, fin de traitement (champs vides à 89,11%) et résultat thérapeutique (champs vides à 54,94%).

Les mesures correctrices que nous avons utilisées pour la complétude des données sont les suivantes : Pour le champ "genre", il a été complété en fonction du prénom du patient. Par contre, pour les autres champs (résultat thérapeutique), nous avons consulté d'autres documents médicaux (demande d'examen d'échantillon biologique pour la tuberculose, carte d'identité du patient tuberculeux, registre des cas de tuberculose).

Les critères de recherche étaient les suivants : *même nom-même prénom-même sexe*, *même nom-même prénom-même sexe-même âge* et *même nom-même prénom-même sexe-même âge-même lieu habitation*. Nous avons observé un total 35 cas (1,69%) de doublons potentiels (patients ayant le même nom et le même prénom). (Figure 34).

Parmi ces doublons potentiels, nous avons dénombré 10 champs vides (c.-à-d.. la valeur répétée 10 fois est la valeur vide) et 25 champs ayant les mêmes noms et prénoms qui

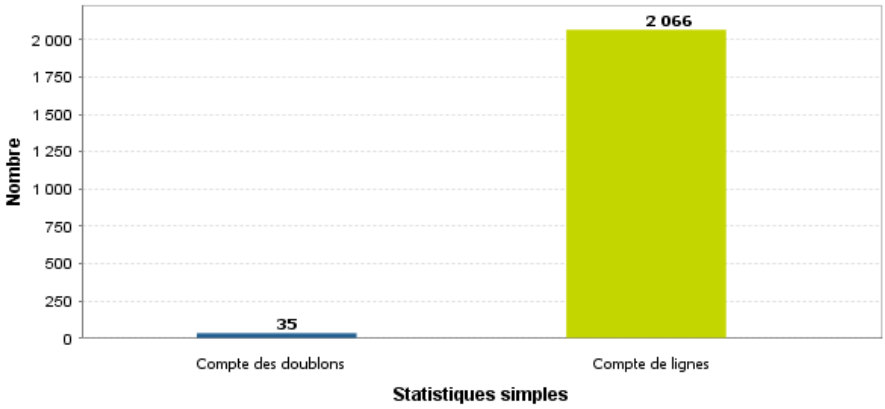


FIGURE 34 – Nombre de doublons potentiels

reviennent 2, 3 et 4 fois. La figure 35 présente la fréquence d'apparition de chaque doublon potentiel détecté. Les noms et prénoms des patients ont été cryptés pour des besoins de confidentialité liés aux données à caractère médical.

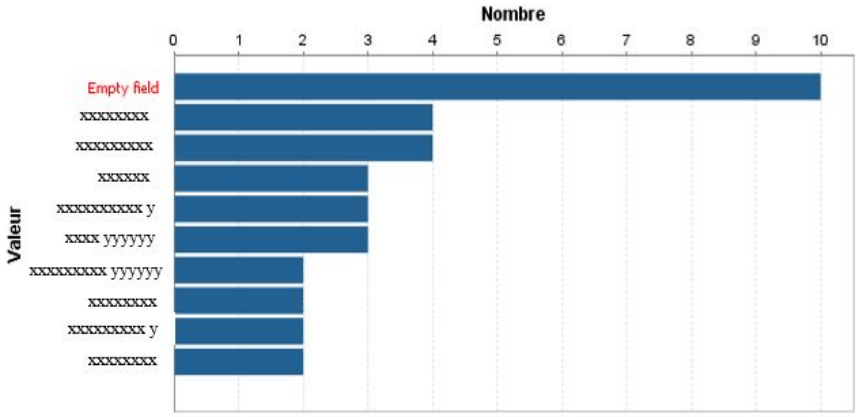


FIGURE 35 – Fréquence d'apparition des doublons potentiels

Les mesures correctrices que nous avons utilisées pour les doublons potentielles sont les suivantes : les patients dont plus de 80% des champs étaient vides ont été supprimés du jeu de données. Par contre, pour les patients dont les mêmes noms et prénoms revenaient 2, 3 et 4 fois, nous avons d'abord contrôlé leur genre, leur âge, leur nationalité et lieu d'habitation. Après contrôle, pour les doublons restants, nous avons gardé un seul enregistrement de chaque patient.

Dimension validité : Pour évaluer la validité des données, nous avons utilisé la méthode de profilage des données par colonne. Le résultat a montré que la variable démographique âge avait 24 valeurs invalides. Certaines de ces valeurs étaient par exemple 3ème âge, 3M, etc.

Pour ces champs, étant donné que les sexagénaires et septuagénaires sont considérés

comme personne du « troisième âge », nous avons considéré pour ces patients l'âge médiane de l'intervalle fermé de 65 ans à 90 ans, soit environ 77 ans.

Données géographiques :

S'agissant des données géographiques, nous avons réalisé différentes opérations afin de les uniformiser :

— Uniformiser le datum et le système de projection local pour le Gabon (Gabon Transverse Mercator) de tous les jeux de données.

— Identifier et corriger les problèmes de limite administratives des quartiers. Par exemple, les limites administratives d'un quartier se trouvant dans un arrondissement auquel il n'appartient pas.

Nous avons réalisé manuellement ces opérations de transformation à l'aide du logiciel SIG QGIS 2.18.

3.1.3.4 Définition de l'architecture SOLAP de la tuberculose et technologies utilisées

Le SOLAP de la tuberculose proposé dans ce travail repose sur une architecture ROLAP. Elle est composée de trois niveaux (i) une plateforme Web, (ii) un entrepôt de données spatiales multidimensionnelles et (iii) un client d'analyses multidimensionnelles (Figure 36). Nous décrivons par la suite chaque niveau de cette architecture.

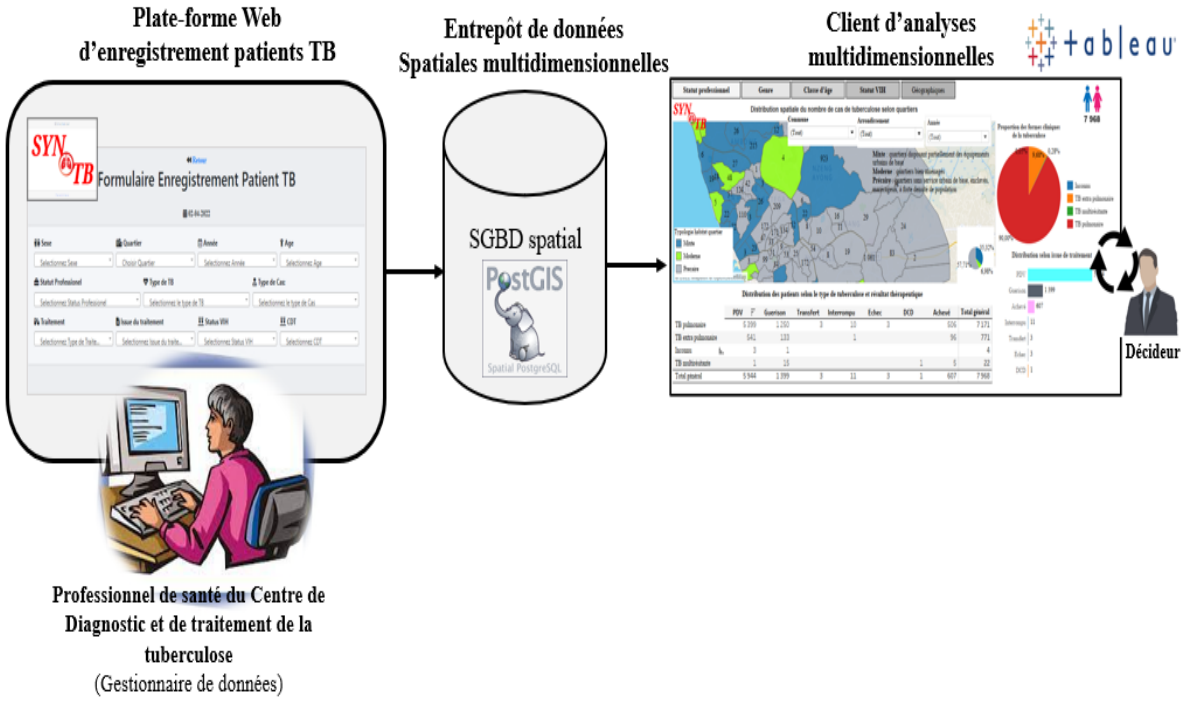


FIGURE 36 – Architecture du prototype SOLAP de la tuberculose proposé

Plateforme Web d'enregistrement des patients TB : A ce niveau, l'ensemble des données cliniques et socio-démographiques des patients sont enregistrées par les gestionnaires des données des centres de diagnostic et de traitement (CDT) de la tuberculose. La plate-forme Web a été développée avec le langage de scripts généraliste et Open Source PHP.

Entrepôt de données : Ce niveau est responsable du stockage des données multidimensionnelles, descriptives (socio-démographiques et cliniques) des patients tuberculeux et géographiques de notre zone d'étude, la région sanitaire Libreville-Owendo-Akanda. L'entrepôt de données a été conçu avec le système de gestion de base de données (SGBD) spatiales PostGIS.

Client d'analyses multidimensionnelles : A ce niveau, les données de l'entrepôt de données spatiales multidimensionnelles sont explorées pour ressortir un ensemble d'indicateurs de suivi et de surveillance de la tuberculose qui sont présentés sous formes de tableaux de bord intégrant de tableaux croisés, graphiques, cartes qui peuvent être visualisés par les experts en santé publique qui sont des décideurs. Nous avons utilisé l'outil d'analyse multidimensionnelle Tableau public [1] qui est une version gratuite de Tableau Software. Car, nous implémentons notre système d'aide à la décision dans un pays à ressources limitées. En plus, cet outil est beaucoup plus facile à utiliser pour les non-informaticiens qui sont en majorité les futurs utilisateurs du système proposé.

3.1.3.5 Modélisation de l'entrepôt de données TB

Méthodologie : Nous avons modélisé les données selon la méthode ROLAP, où les faits et les dimensions sont traduits au niveau logique sous la forme de relations (Bimonte et al., 2016b).

3.1.3.5.1 Résultat : Dans le modèle multidimensionnel proposé, les données ont été stockées selon un schéma en étoile (Figure 37). Le fait étudié est le «*traitement patient tuberculeux*» avec pour mesure le «*nombre de patients*». Il a pour dimensions *Statut VIH, Type_Patient, Genre, Age, Lieu résidence, Résultat thérapeutique, CDT, Type tuberculose, Statut Professionnel et Temps*. Il faut relever que dans ce modèle, la dimension *traitement proposé* n'a pas été prise en compte. Car, notre étude n'avait pas pour objectif de réaliser une étude de survie [2]. Ci-après, nous décrivons chacune des dimensions ou axes d'analyse.

1. https ://www.tableau.com
2. Une étude de survie sert à calculer la variation, au cours du temps, de la probabilité de survie de patients à partir d'un instant t0. On intègre donc progressivement des nouveaux patients dans l'étude et, à certaines dates, on compte le nombre de personnes encore en vie, décédées à cause de la pathologie étudiée ou disparues de l'étude. (https ://www.adscience.fr/uploads/ckfiles/files/html$_{f}iles/StatEL/statel_{c}ourbe_{s}urvie.htm$).

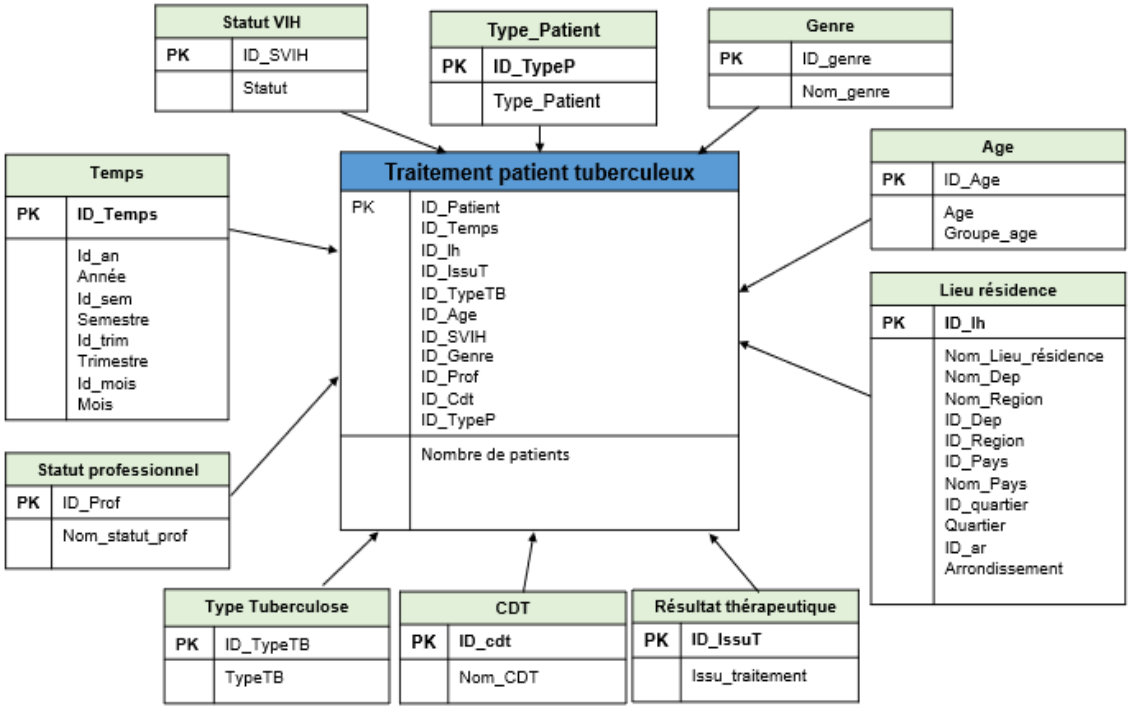


FIGURE 37 – Modèle d'implantation du cube traitement patient tuberculeux (en étoile)

— Statut VIH : le VIH étant une maladie opportuniste, un examen biologique doit être réalisé aux patients diagnostiqués TB positif, afin de voir si sa tuberculose est liée au VIH ou non. Le statut VIH d'un patient TB peut être positif, négatif ou inconnu dans le cas où celui-ci n'est pas renseigné dans son dossier médical.
— Genre : Il s'agit du genre du patient (masculin ou féminin) tuberculeux.
— Âge : l'âge du patient que nous avons regroupé par tranche en tenant compte des tranches d'âge utilisées par le PNLT. À savoir 0-4 ans, 5-9 ans, 10-14 ans, 15-19 ans, 20-24 ans, 25-29 ans, 30-34 ans, 35-39 ans, 40-44 ans, 45-49 ans, 50-54 ans, 55-59 ans, 60-64 ans, 65-69 ans, 70-74 ans, 75-79 ans et 80 ans et plus
— Lieu résidence : Il s'agit du quartier (identifié par son nom) où la personne diagnostiquée tuberculose positive réside.
— Résultat thérapeutique : Il s'agit de l'issue du traitement d'un patient mis sous anti-tuberculeux (guérison, échec de traitement ou thérapeutique, traitement interrompu, traitement terminé ou achevé, décès, perdu de vue, non évalué).
— CDT : le centre de diagnostic et de traitement de la tuberculose dans lequel le patient a été pris en charge. Les CDT sont identifiés à partir de leur nom (par exemple : CDT Nkembo).
— Type de tuberculose : la forme clinique de la tuberculose selon la localisation (pulmonaire ou extra-pulmonaire). Mais aussi la résistance aux antituberculeux (TB- multirésistante ou TBMR).

— Type de patient : le patient tuberculeux peut être un nouveau cas, un ancien cas ou cas de rechute.
— Statut professionnel : le statut professionnel du patient, regroupé en fonctionnaires, artisans, commerçants, élèves-étudiants, employés-ouvriers, professionnels de la santé, retraités, sans emplois ou inconnu.
— Temps : il s'agit de la dimension temporelle structurée selon quatre niveaux hiérarchiques (année, semestre, trimestre et mois).

3.1.3.6 Création et chargement de l'entrepôt de données

Nous avons conçu l'entrepôt de données de la tuberculose en utilisant le SGBD spatial PostGIS. Après création des différentes tables (fait et dimensions), nous avons peuplé les tables de dimensions descriptives et la table de fait avec les données réelles collectées. Ce peuplement s'est fait par importation des fichiers CSV de chacune des dimensions et de la table de fait. Ensuite, nous avons peuplé les dimensions géographiques (quartiers, arrondissements et communes) en important les fichiers Shapefile de quartiers, arrondissement et communes dans les tables de dimensions correspondantes. Le gestionnaire de base de données DB Manager du logiciel SIG QGIS 2.18 a été utilisé pour cette opération.

3.1.3.7 Conception et validation tableaux de bord

Méthodologie : Pour l'élaboration des tableaux de bord, nous avons utilisé l'outil de l'informatique décisionnel ou BI Tableau comme client d'analyse multidimensionnelle. Dans un premier temps, nous avons procédé avec les responsables du PNLT à la sélection et à la validation d'une liste d'indicateurs qu'ils souhaitent suivre et surveiller sous forme de tableaux de bord :

— Nombre de cas de TB par année ;
— Pourcentage des cas de TB selon le statut professionnel ;
— Pourcentage des cas de TB selon la forme clinique ;
— Pourcentage des cas de TB selon le statut VIH ;
— Nombre de cas de TB selon la classe d'âge ;
— Nombre de cas de TB selon le quartier ;
— Pourcentage des cas de TB selon le quartier ;
— Nombre de cas de TB selon l'issue thérapeutique et la forme clinique de la TB

Cette liste a été par la suite complétée par d'autres indicateurs, tels la prévalence de la tuberculose par quartier et la prévalence de tuberculose par arrondissement.

Dans un second temps, nous avons réalisé plusieurs analyses de données (univariée et bivariée) en explorant l'entrepôt de données spatiales multidimensionnelles selon les différentes dimensions. En outre, les résultats d'analyses ainsi que les différents modes de visualisations d'indicateurs ont été discutés et validés avec les responsables du PNLT.

Enfin, nous avons conçu les différents tableaux de bord dynamiques et interactifs en tenant compte des indicateurs de suivi proposés par le PNLT et les axes d'analyse.

Résultat : La figure 38 présente les différents tableaux de bord. Dans ces tableaux de bord, il est possible pour l'utilisateur de sélectionner l'information en appliquant des filtres qui tiennent compte des dimensions géographiques et temporelles. Aussi, ces tableaux de bord ont été validés par les responsables du PNLT lors de séances de présentations. Les différentes visualisations ont été par la suite utilisées pour la production d'une narration des données sur la tuberculose au Gabon (Section 3.2.2).

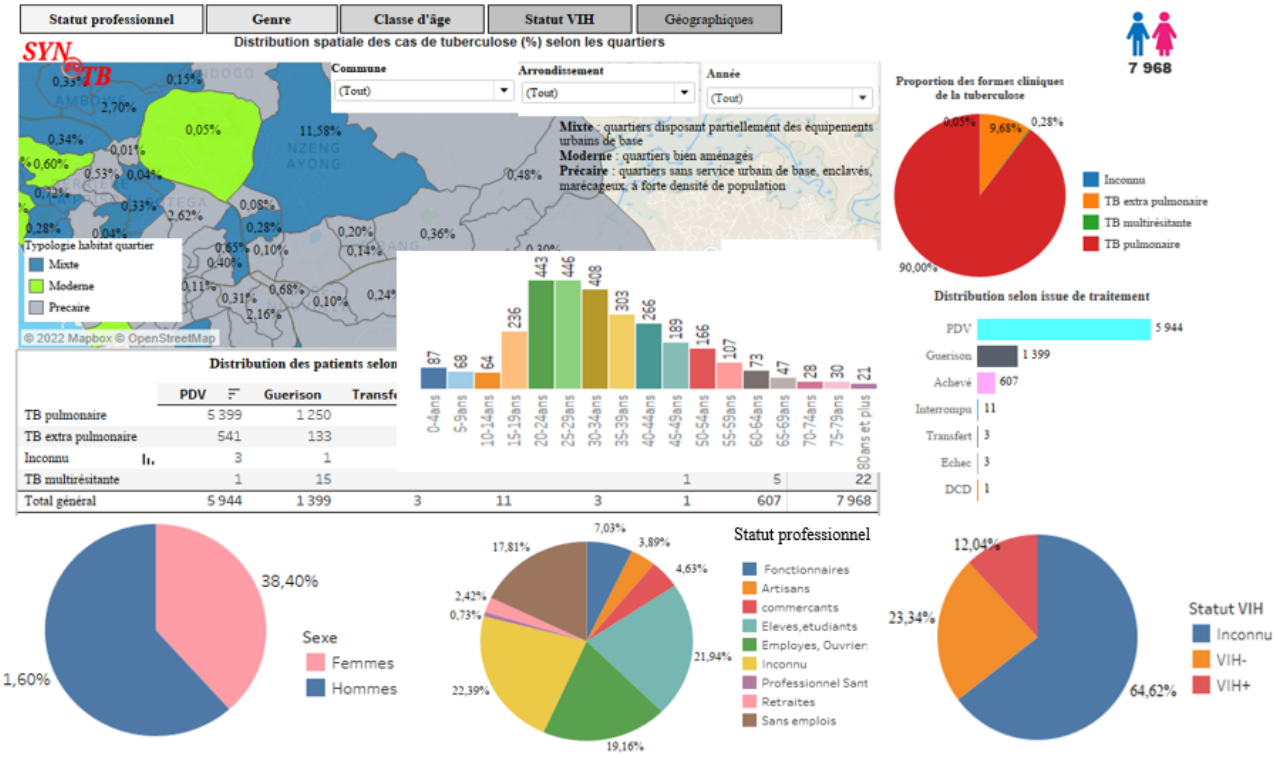


FIGURE 38 – Tableaux de bord : visualisation écran d'accueil et selon les différents axes d'analyse (statut professionnel, statut VIH, type de tuberculose, résultat thérapeutique, géographique, sexe et âge).

3.1.3.8 Conception et validation plateforme Web d'enregistrement des cas TB

Contexte : A ce jour, le Gabon ne dispose pas de moyen numérique pour l'enregistrement des personnes diagnostiquées positives à la tuberculose. Le système d'enregistrement des patients tuberculeux reste encore manuel à tous les niveaux de la pyramide sanitaire.

Objectif : La plate-forme Web que nous avons conçu dans la cadre de ce travail a pour objectif d'améliorer la promptitude et complétude des données par l'enregistrement

en quasi-temps réel des patients tuberculeux au niveau des centres de diagnostic et de traitement de la tuberculose (CDT).

Méthodologie : La plateforme Web [3] d'enregistrement des patients tuberculeux a été développée avec le langage de script PHP avec pour SGBD PostgreSQL avec son extension spatiale PostGIS (Section 3.1.3.4).

Résultat : L'accès à cette plateforme peut se faire avec une session administrateur ou gestionnaire de données.

Interface administrateur : L'administrateur est un agent du service Suivi-Évaluation des activités de lutte du PNLT. Il peut réaliser les tâches suivantes (Figure 39) : (i) créer un gestionnaire de données en l'affectant à un CDT, (ii) modifier les informations d'un gestionnaire de données, (iii) supprimer un gestionnaire de données, (iv) créer un nouveau CDT, (v) modifier les informations d'un CDT, (vi) supprimer un CDT et (vii) consulter la liste des cas TB enregistrés par années.

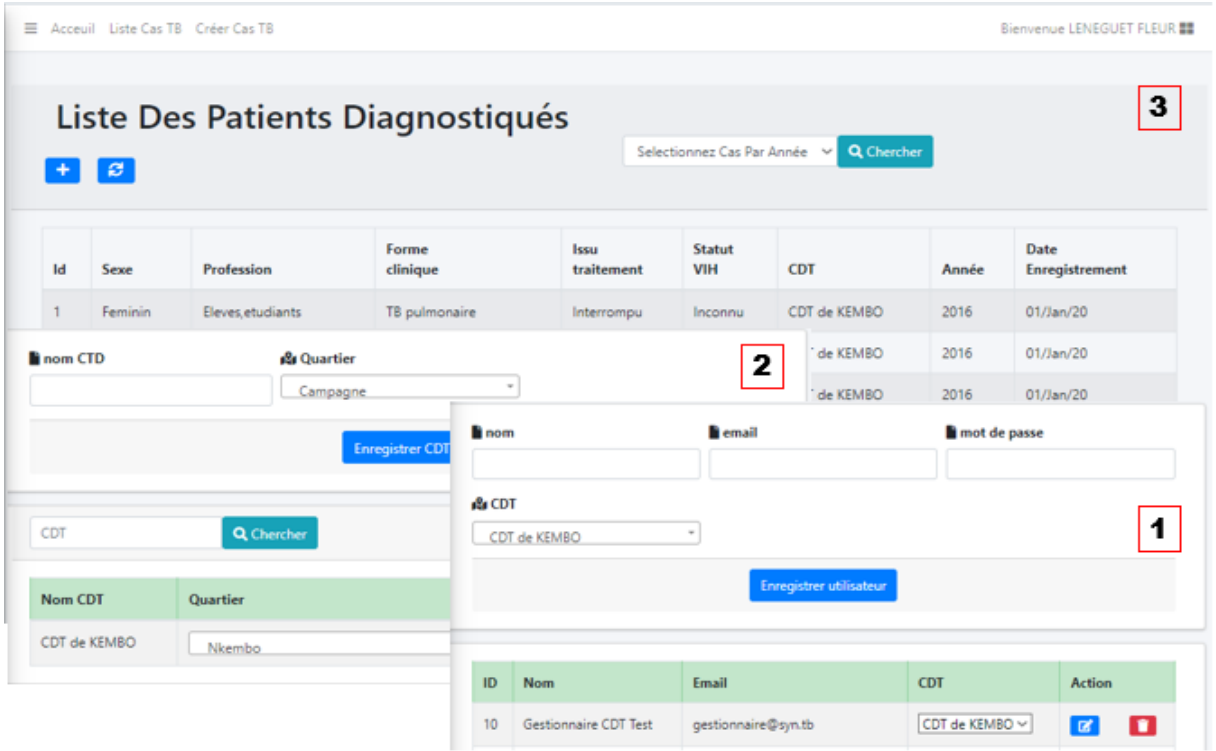


FIGURE 39 – Plate-forme Web : tâches administrateur (1= enregistrement des gestionnaires données ; 2= enregistrement des centres de diagnostic et de traitement (CDT) et 3=Visualisation des patients tuberculeux enregistrés

3. https ://syntb-egabonsis.info

Interface gestionnaire de données : Un gestionnaire de données est un agent d'un centre de diagnostic et de traitement de la tuberculose (CDT) chargé de la collecte de données afin de renseigner le système de surveillance épidémiologique de la tuberculose.

Dans cette interface, les tâches suivantes sont réalisables (Figure 40) : (i) enregistrer les patients tuberculeux de son CDT via un formulaire qui prend en compte les caractéristiques socio-démographiques et cliniques des patients et (ii) consulter la liste des patients en faisant une recherche par année. L'ensemble des données enregistrées par les gestionnaires de données des CDT sont stockées dans l'entrepôt de données spatiales multidimensionnelles (EDSM) de la tuberculose sous PostGIS.

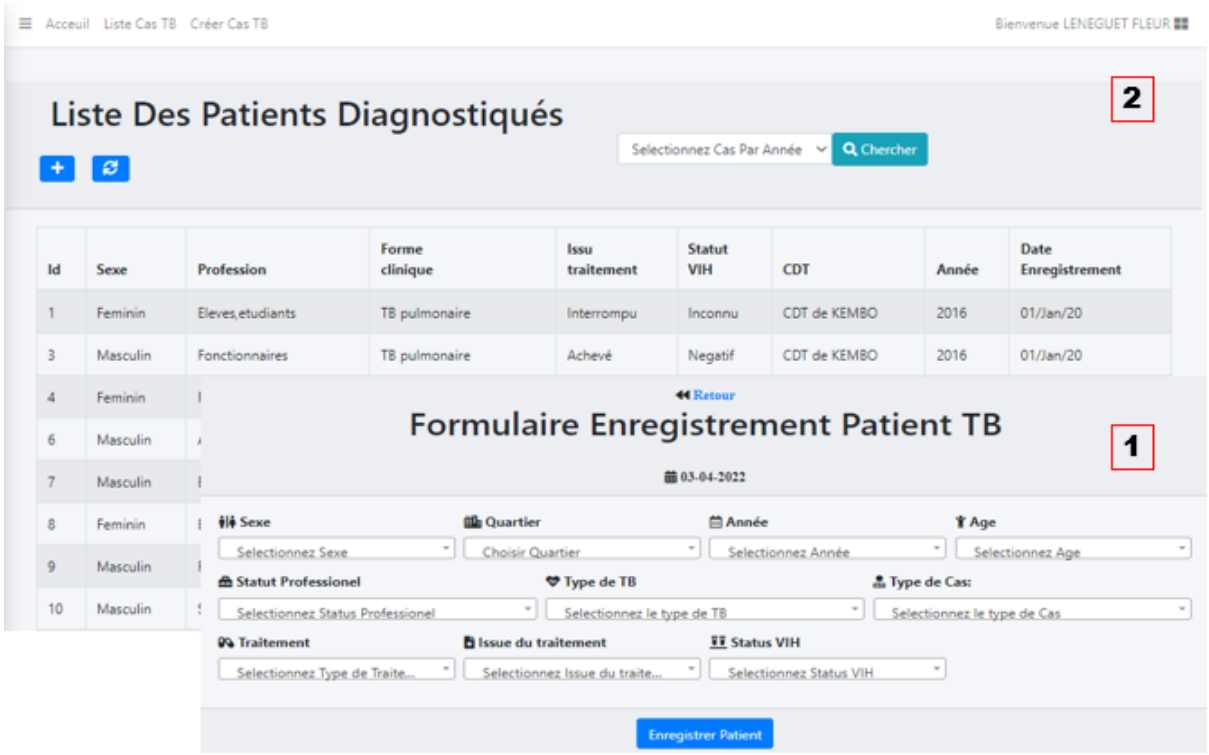


FIGURE 40 – Plate-forme Web : tâches gestionnaires de données

3.1.3.9 Formation et évaluation du SOLAP-TB

Nous avons enfin formé les utilisateurs (responsables du PNLT et gestionnaires de données) à l'utilisation du système d'aide à la décision spatio-temporel de la tuberculose. Ensuite, son utilisation par les responsables du PNLT et les gestionnaires de données des CDT a été évaluée.

Méthodologie : Deux questionnaires comportant des questions ouvertes et fermées ont été conçus. Les questions concernaient l'utilisation de la plateforme Web et du Tableau de Bord pour l'aide à la décision.

Résultats : Dans la population enquêtée, il y avait trois responsables du PNLT et quatre gestionnaires de données de CDT.

Concernant l'évaluation de l'utilisation par les responsables du PNLT, tous (100%) ont répondu que :

- R1 : la plate-forme Web permettra d'améliorer l'actuel système de notification des cas déclarés tuberculeux dans les centres de diagnostic et de traitement (CDT) ;
- R2 : la plate-forme Web permettra d'améliorer la complétude des données du programme nationale de lutte contre la tuberculose ;
- R3 : le programme national de lutte contre la tuberculose du Gabon devrait adopter cette plate-forme Web, comme plateforme nationale de notification des cas déclarés TB par les CDT ;
- R4 : les différents tableaux de bord prennent en compte les indicateurs utiles au suivi de l'évolution de l'épidémie de tuberculose ;
- R5 : À partir des résultats d'analyse présentés dans les différents tableaux de bord, ils vont pouvoir mieux orienter leurs actions de ripostes contre la tuberculose.

Par rapport à la question R5, un des responsables du PNLT a déclaré : *« Parce que les résultats d'analyse permettent de cibler les aspects sur lesquels il faut agir afin d'orienter et de mettre en œuvre des actions précises en matière de riposte contre la TB»*. Il ajoute : *« ces résultats permettent de mener des actions visant à réduire le nombre de Perdus de vue et orienter nos interventions selon les zones spatiales les plus touchées »* .

Les responsables ont proposé quelques actions qui peuvent être menées :

1. L'extension des Centres de Traitement (CT) dans les quartiers qui ont un nombre de cas de TB très important, afin de désengorger le CDT de Nkembo qui reçoit plus de la moitié des cas de TB du Gabon.
2. Focaliser une activité précise de dépistage de masse dans une région sanitaire présentant une incidence de la TB pédiatrique élevée.

Les trois responsables (100%) ont répondu être satisfaits par rapport au rendu des différents tableaux de bord en termes d'indicateurs de surveillance épidémiologique de la tuberculose.

Pour ce qui est de l'évaluation de l'utilisation de plate-forme Web par les gestionnaires de données des CDT, nous avons recueilli les réponses suivantes :

- 100% ont répondu qu'en dehors d'un éventuel problème d'accès à un Internet, avec leur identifiant et mot de passe, ils n'ont eu aucune difficulté à accéder à la plateforme Web pour l'enregistrement des patients tuberculeux.
- 75% des gestionnaires de données trouvent le formulaire d'enregistrement des cas de tuberculose facile à renseigner et 25% très facile à renseigner.
- Pour 50% des gestionnaires de données, l'affichage de la liste du nombre de cas déclarés tuberculeux par année est facile à afficher. Par contre, 25% ne la trouve pas du tout facile à afficher et 25% très facile à afficher.

— 100% ont répondu que la plate-forme Web permettra d'améliorer l'actuel système de notification des cas déclarés tuberculeux dans les centres de diagnostic et de traitement (CDT).
— 100% ont répondu que la plate-forme Web permettra d'améliorer la complétude des données du programme nationale de lutte contre la tuberculose.
— 100% ont répondu que le programme national de lutte contre la tuberculose du Gabon devrait adopter cette plate-forme Web comme plateforme nationale de notification des cas déclarés TB par les CDT.

3.1.4 Discussion

Dans le processus de conception et de réalisation du système d'information décisionnel de la tuberculose, plusieurs problèmes se sont posés à nous. Parmi ceux-ci, il y a d'abord celui de la qualité des données de l'infection tuberculeuse, comme le montre les résultats de l'analyse de la qualité des données (sous-section 3.1.3.2). Ce problème de la qualité des données est également retrouvé dans de nombreux systèmes de surveillance de maladies (Beyeme-Ondoua, 2007; Palussière et al., 2013; Majerovich et al., 2017) que nous avons pu étudier. Ensuite, au niveau du modèle conceptuel des données, la situation médicale (par exemple tuberculose ou VIH positif) de certains statuts professionnels (militaire, gendarme, policier) ne devant pas être dévoilée, pour des raisons stratégiques, à la dimension « statut professionnel », nous avons procédé à des regroupements de statuts professionnels. Ces regroupements ne permettent pas de réaliser une analyse plus fine afin de connaître de façon détaillée les statuts professionnels plus impactés par l'infection tuberculeuse au Gabon. Aussi, notre étude n'étant pas une étude de survie, nous n'avons pas tenu compte des traitements administrés aux patients tuberculeux dans le modèle conceptuel. Pour ce qui est de la dimension statut VIH, nous ne pouvons pas également dire grand-chose dans le cadre de la co-infection tuberculose-VIH. Car, sur les 7 968 dossiers médicaux des patients que nous avons consultés pour la collecte des données, cette variable n'est quasiment pas renseignée. Or, la tuberculose est une des principales causes de mortalité des personnes infectées par le virus de l'immunodéficience humaine (VIH) (Straetemans et al., 2010). Enfin, la non-identification des patients infectés par la tuberculose de façon unique dans le système de surveillance est source de doublons. Car, un patient tuberculeux qui se déplace par exemple pour des raisons économiques d'une ville à une autre peut s'enregistrer de nombreuses fois comme nouveau cas dans différents centres de diagnostic et de traitement de la tuberculose. Dans ce cas, il est compté en double dans le nombre de nouveaux cas de tuberculose. Ce qui biaise les données statistiques liées à cette maladie et entraîne par conséquent des mauvaises prises de décisions dans la stratégie de riposte antituberculeuse.

3.1.5 Conclusion

Dans cette section, nous avons présenté de façon détaillée le système d'aide à la décision spatio-temporel de la tuberculose. Ce système repose sur une architecture à trois niveaux, une plate-forme Web d'enregistrement des personnes souffrant de tuberculose, un entrepôt de données spatiales multidimensionnelles et un client d'analyse multidimensionnelle. Les objectifs visés par ce système sont d'améliorer (i) la notification des cas de tuberculose en permettant aux centres de diagnostic de traitement de la tuberculose d'enregistrer en quasi-temps réel les personnes tuberculose positive via une plate-forme Web et ainsi diminuer le problème de qualité des données retrouvé dans l'actuel système de surveillance, (ii) le stockage et l'accès aux données TB via l'entrepôt de données spatiales multidimensionnelles, pour des besoins spécifiques (par exemple, la recherche, les enquêtes, etc.) et (iii) la prise de décisions stratégiques grâce à la production de tableaux de bord dynamiques et interactifs d'indicateurs spatio-temporels de la tuberculose.

Le SOLAP de la tuberculose proposé permet une analyse multidimensionnelle des données hétérogènes impactant sur la propagation de la tuberculose. Aussi, il est facilement adaptable pour la surveillance épidémiologique de la tuberculose pour d'autres régions ou pays. Par contre, du fait de la spécificité des indicateurs de la tuberculose, il n'est pas adaptable pour la surveillance d'autres maladies. Mais le processus suivi pour sa conception et sa réalisation peut être réutilisé pour la conception d'un système décisionnel pour la surveillance d'une autre maladie.

En perspective, après la phase d'implémentation dans notre région d'étude (Libreville-Owendo-Akanda), nous proposons la prise en compte des neuf régions sanitaires restantes. Cela permettra aux experts de la santé qui sont décideurs d'avoir en quasi-temps réel une vision globale de situation épidémiologique de la maladie au Gabon. Aussi, une prise en compte de la dimension « traitement administré » dans le modèle conceptuel est nécessaire. Cela permettra d'analyser dans l'espace et dans le temps les différents traitements administrés aux patients en termes de molécules. Ainsi, en croisant par exemple ces traitements avec les résultats thérapeutiques, cela permettra aux épidémiologistes et médecins de mieux suivre les succès thérapeutiques (guérison) par rapport aux traitements administrés aux patients.

Enfin, une identification unique du patient tuberculeux permettra de limiter les doubles comptes dans le nombre de personnes réellement infectées par la tuberculose au Gabon.

3.2 Processus de narration de données en intelligence épidémique proposé

Contexte : Le Gabon, à l'instar d'autres pays d'Afrique, est confronté à la persistance de certaines maladies comme la tuberculose. Mais, des moyens classiques sont encore uti-

lisés pour sensibiliser et éduquer la population (par exemple : affiches, flyers) et pour présenter la situation épidémiologique (par exemple : rapports d'activité, bulletin épidémiologique) aux autorités sanitaires. Un des moyens les plus efficaces utilisé actuellement pour communiquer au public les faits importants dans un domaine (par exemple : santé publique, économie, politique, agriculture) est la narration des données. Dans le domaine de la santé publique, les narrations de données retrouvées dans la littérature visent majoritairement le grand public pour la sensibilisation sur les problèmes de santé publique et l'éducation thérapeutique des patients. Aussi, ces narrations de données ont été élaborées selon un processus ad hoc, c'est-à-dire qu'ils ne tiennent pas compte des bonnes pratiques des processus généraux de narration de données et du processus d'intelligence épidémique de l'Organisation Mondiale de la Santé.

Objectif : Notre objectif est de proposer un processus spécifique de narration de données pour l'intelligence épidémique. Ce processus devra satisfaire un certain nombre d'exigences, à savoir : (i) personnaliser les processus généraux de narration de données en intégrant les caractéristiques spécifiques d'un processus d'intelligence épidémique de l'OMS, (ii) transmettre des résultats scientifiques à un public d'experts afin de rendre compte de la situation sanitaires et des résultats des politiques de santé mises en œuvre et, plus généralement, d'aider à la prise de décision et (iii) être générique. Cet objectif représente notre deuxième verrou scientifique baptisé processus de narration de données en intelligence épidémique (P-N-D-I-E).

À la section 3.2.1, nous décrivons comment le processus de narration de données en intelligence épidémique a été conçu, les différentes phases ainsi que les activités. Ensuite, à la section 3.2.2, nous illustrons l'utilisation de ce processus par la production d'une narration de données sur la tuberculose au Gabon.

3.2.1 Conception du processus de narration de données en intelligence épidémique

Méthodologie : L'étude de l'état de l'art a permis de ressortir les phases des processus généraux de narration de données sur lesquelles s'accordent de nombreux auteurs (Lee et al., 2015; Chen et al., 2020; El Outa et al., 2022) (Figure 41.A) et les phases du processus d'intelligence épidémique (Figure 41.B) recommandé par l'Organisation Mondiale de la Santé (WHO, 2014). Une analyse comparative de ces deux processus a montré que les quatre premières phases du processus d'intelligence épidémique de l'OMS (détection, triage, vérification et évaluation du risque épidémique) (Figure 41.B) correspondent à celle d'exploration des données des processus généraux de narration de données (Figure 41.A). Tandis que la phase communication (Figure 41.B) correspond à celle de présentation (Figure 41.A).

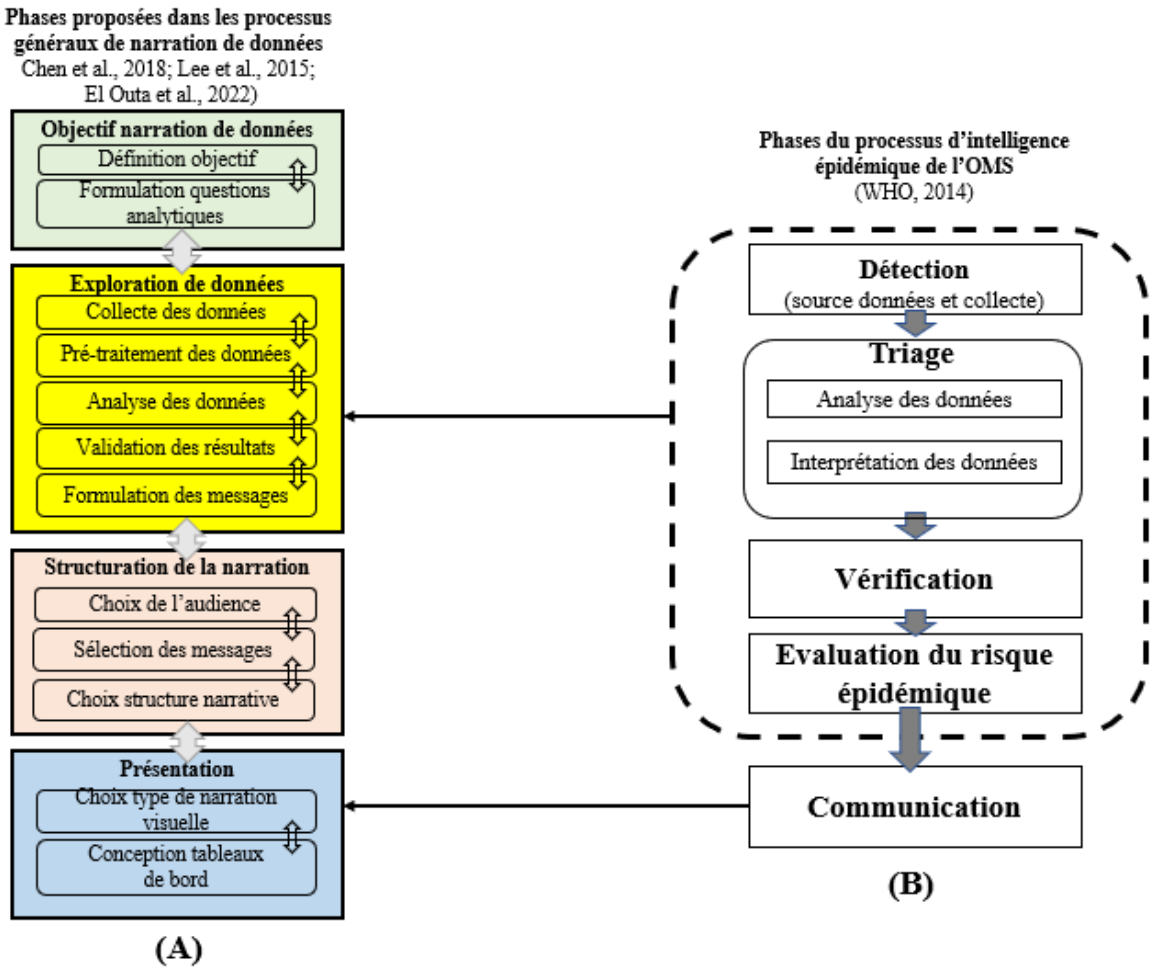


FIGURE 41 – Phases des processus généraux de narration de données (A) et du processus d'intelligence épidémique de l'OMS (B)

À partir de ces deux processus, nous avons proposé un processus de narration de données en intelligence épidémique (Figure 42) qui s'inspire des processus généraux de narration de données (phases et activités) et est enrichi avec les activités particulières du processus d'intelligence épidémique (vérification et évaluation du risque épidémique). Ces deux activités du processus de l'OMS sont essentielles dans le processus de surveillance épidémiologique des maladies. En effet, l'activité de vérification a pour objectif de comparer les résultats de l'analyse des données à l'état de l'art (par exemple, comparaison à un seuil défini de l'Organisation Mondiale de la Santé, situation dans d'autres régions, pays, continents, etc.) afin de mieux évaluer le risque d'épidémie.

Résultat : Le processus proposé dans ce travail se décompose en quatre phases (Figure 42) (objectif de la narration de données, exploration des données, structuration de la narration et présentation). À chaque phase, un ensemble d'activités sont réalisées en suivant une séquence logique. Toutefois, en cas de besoin, il est possible de faire des allers-retours entre phases ou activités. Par exemple, à la phase d'exploration des données, le produc-

teur de la narration de données peut retrouver des résultats devant le ramener à la phase objectif afin formuler une ou des nouvelles questions analytiques. Il est également possible de revenir à une activité précédente. Par exemple, à l'activité d'analyse des données de la phase exploration des données, le producteur de la narration de données peut trouver des résultats aberrants (par exemple : nombre de cas de tuberculose plus élevé chez les enfants de 0-11 mois dans la population) devant le contraindre à réaliser une nouvelle activité de collecte de données.

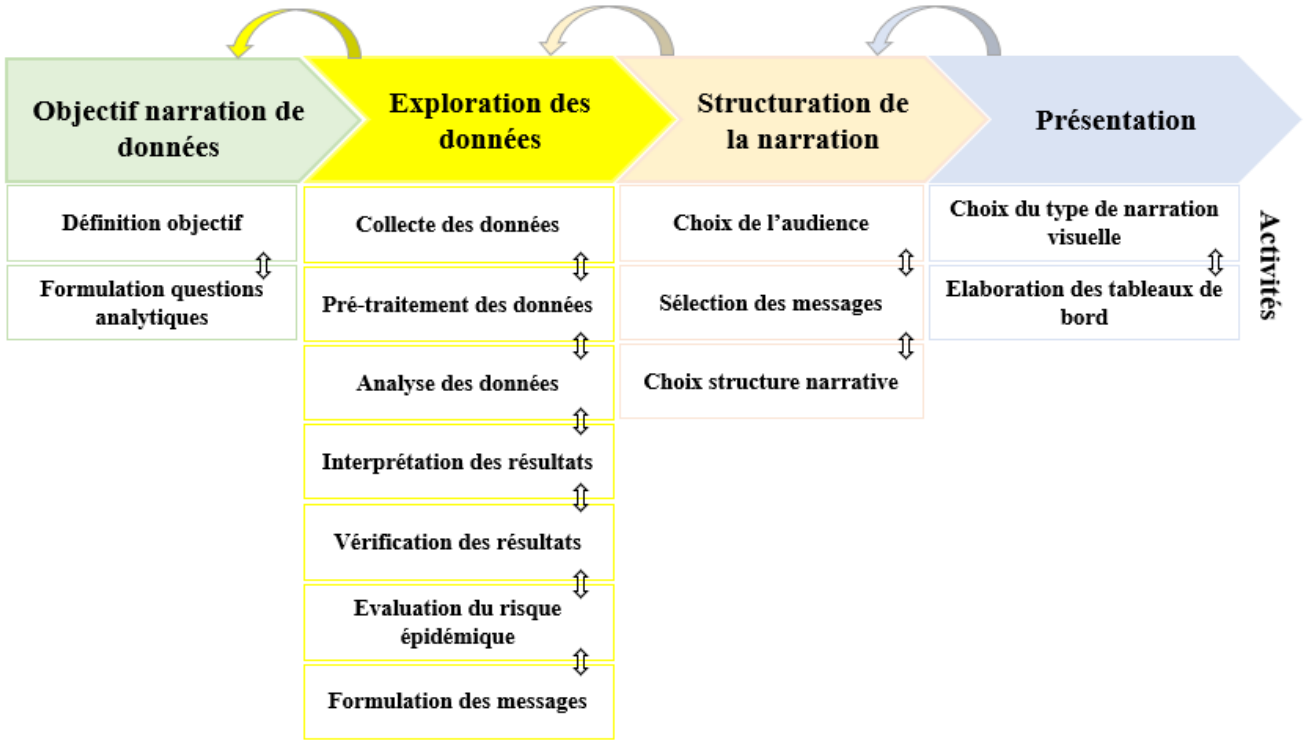


FIGURE 42 – Processus de narration de données en intelligence épidémique proposé

Dans les sous-sections ci-après, nous décrivons de façon détaillée les activités de chacune des phases du processus.

3.2.1.1 Objectif de la narration de données

À cette première phase, deux (2) activités sont réalisées : la définition de l'objectif de la narration de données et la formulation des questions analytiques.

— *Définition de l'objectif* : cette activité consiste à définir clairement l'objectif visé par la production de la narration de données. Il s'agit d'une étape intentionnelle, qui guidera la phase d'exploration des données. Par exemple, dans le domaine de la santé publique, une narration peut avoir pour objectif de décrire le profil épidémiologique d'une maladie.

— *Formulation des questions analytiques* : une fois l'objectif de la narration de données défini, il doit être décomposé en questions analytiques (par exemple, quels sont les

facteurs qui favorisent la persistance d'une maladie données ? quelles sont les zones géographiques à risque pour cette maladie ?, etc.). Ces questions servent à guider le producteur de la narration de données dans la phase d'exploration des données. Par exemple, lors de l'activité de collecte de données, ces questions analytiques lui permettent de procéder à une sélection des sources de données pouvant permettre de répondre à ces différentes questions. Ces questions sont également utiles pour l'activité analyse des données.

3.2.1.2 Exploration des données

Après la définition de l'objectif de la narration avec les différentes questions analytiques qui en découlent, vient la phase d'exploration des données. À cette phase, sept (7) activités sont réalisées :

- *La collecte des données* : cette activité consiste à collecter toutes les données (sociodémographiques, cliniques, géographiques, etc.) utiles pour la production de la narration de données. Il s'agit des données qui seront analysées pour répondre aux questions analytiques posées à la première phase du processus.
- *Le prétraitement des données* : les données collectées peuvent présenter plusieurs problèmes de qualité (champs partiellement remplis ou vides, doublons, etc.). L'activité traitement des données consiste à gérer les données manquantes, à corriger les incohérences, à supprimer les doublons et plus généralement à résoudre des problèmes de qualité. À cette activité, un outil de gestion de la qualité des données (par exemple Talend data quality) sera utilisé.
- *L'analyse des données* : elle consiste en une épidémiologie descriptive et analytique, c'est-à-dire l'organisation des données en temps, lieu et facteurs de risque (par exemple, les personnes les plus à risque de développer la tuberculose sont celles dont le système immunitaire est affaibli : personnes VIH Positif, personnes âgées) ou d'exposition (par exemple l'historique des voyages, lieux fréquentés par une personne infectée). À cette activité, un logiciel d'analyse statistique (par exemple Rs Studio, SPSS, Epi Info) ou un outil d'analyse et de visualisation des données (par exemple Tableau Software ou Power BI) sera utilisé pour la production de tableaux, de graphiques, de cartes, etc. Cette activité permet de faire ressortir des trouvailles qui sont des signaux (information/donnée traitée - triée et identifiée comme contenant des informations en relation avec un risque sanitaire) correspondant à des messages intéressants à communiquer.
- *L'interprétation des résultats* : cette activité consiste à donner un sens particulier aux trouvailles qui sont ressorties de l'activité d'analyse de données.
- *La vérification des résultats* : cette activité consiste à comparer les trouvailles à ceux d'autres travaux sur la maladie étudiée. Mais aussi à ceux d'autres sources fiables telles que les instituts de santé publique ou les laboratoires. L'objectif est

de voir si ces trouvailles sont confirmés par la littérature ou s'ils apportent de nouvelles connaissances par rapport au problème de santé publique étudié.

— *L'évaluation du risque épidémique* : Cette activité consiste à caractériser les trouvailles et d'estimer l'importance du risque qu'elles représentent pour la santé publique. Lorsqu'une trouvaille a été évaluée comme une menace pour la santé publique, une alerte de risque de propagation régionale ou nationale est lancée. Les critères d'évaluation du risque épidémique sont par exemple le nombre de cas, les décès, le mode de transmission, les multiples foyers, l'échec thérapeutique et la résistance aux anti-infectieux. Par exemple, lors de l'activité d'analyse des données, en trouvant un très grand nombre de cas de tuberculose multirésistante (TBMR) et de perdus de vue, si ces cas sont vérifiés, alors cela peut être considéré comme un risque pour la santé publique. Car, les personnes ayant développé une TBMR demeure infectieux plus longtemps, ce qui accroît le risque d'exposition de la population. Une fois le risque épidémique évalué, les autorités sanitaires doivent mettre en place une politique sanitaire de riposte adéquate afin de freiner la propagation de la maladie. Par exemple, pour le cas des perdus TBMR, ils peuvent demander au programme de lutte contre la tuberculose de lancer une campagne de recherche active de ces cas.

— *La formulation des messages* : A cette activité, les messages clés à communiquer aux décideurs sont définis. Ils doivent permettre d'attirer leur attention sur la situation épidémiologique afin qu'ils déclenchent rapidement des actions de lutte.

3.2.1.3 Structuration de la narration

À cette phase, trois (3) activités sont réalisées :

— *Choix de l'audience* : une narration de données peut s'adresser à des publics différents (par exemple : grand public, décideurs politiques, professionnels de la santé, chercheurs, experts, etc.). Selon les domaines, il est donc important de bien cibler le public à qui s'adresse la narration de données. Par exemple, dans le domaine de la santé publique, les messages de sensibilisation conviennent mieux au grand public. En revanche, les tendances en matière d'incidence et de létalité liées à une maladie intéresseront davantage les experts en santé publique qui sont des décideurs.

— *Sélection des messages* : Parmi les messages formulés à l'activité *formulation de messages* de la phase l'exploration des données, les messages clés à communiquer au public cible sont choisis. L'objectif visé est de transmettre les messages les plus importants sur la situation épidémiologique de la maladie.

— *Choix de la structure narrative* : Enfin, les messages clés à communiquer sont ordonnés et organisés de manière cohérente et logique, adaptés à l'audience. Plus précisément, l'intrigue du récit est modélisée comme un ensemble d'actes, regroupant des messages à relayer. Aussi, la manière de naviguer dans les actes est également

définie. Par exemple, le premier acte de la narration peut introduire le récit. Ensuite, au deuxième, présenter des messages importants. Enfin, à partir ce deuxième acte, accéder aux autres actes. Ce type de structure porte le nom de Martini Glass (Segel and Heer, 2010).

3.2.1.4 Présentation

À cette phase, il faut procéder au choix de visualisations pour mieux faire passer les messages et capter l'attention de l'audience. Chaque acte est représenté par une unité de visualisation, appelée tableau de bord. Au sein des actes, chaque message est représenté par un ou plusieurs artefacts visuels (graphiques, cartes, tableaux, texte, images, audio). Deux (2) activités sont réalisées à cette phase :

— *Choix du type de narration visuelle* : A cette activité, on procède au choix du moyen (exemple : PowerPoint, tableaux de bord interactifs, vidéo, etc.) qui sera utilisé pour le rendu de la narration de données.

— *Élaboration des tableaux de bord* : A cette activité, il faut tester toutes les possibilités de visualisation et mettre en place les différents graphiques, cartes et tableaux en accord avec les futurs utilisateurs (décideurs). Chaque acte est représenté par un tableau de bord interactif. Tandis que chaque épisode (concernant un message) est représenté par un ou plusieurs artefacts visuels (par exemple, des graphiques, des cartes, des tableaux, du texte, des images, du son).

À la section 3.2.2, nous présentons un exemple pratique d'utilisation du processus proposé à travers la production d'une narration de données sur la pandémie de la tuberculose au Gabon (Ondzigue Mbenga et al., 2022).

3.2.2 Une narration de données sur la tuberculose au Gabon

Dans cette section, nous présentons les phases et activités du processus de production d'une narration de données sur la pandémie de la tuberculose au Gabon. Cette narration de données s'appuie sur le système d'aide à la décision spatio-temporel (SOLAP-TB) proposé (cf. section 3.1). Les sous-sections ci-après présentent les différentes phases et activités pour la production de cette narration de données.

3.2.2.1 Définition de l'objectif de la narration de données

Dans cette narration de données, nous décrivons le profil épidémiologique de la pandémie de la tuberculose dans la région sanitaire Libreville-Owendo-Akanda du Gabon. Pour pouvoir répondre à cet objectif, nous avons mené des entretiens auprès de divers responsables du programme national de lutte contre la tuberculose (directeur, responsable suivi-évaluation, responsable de la prise en charge) et médecins impliqués dans la lutte

antituberculeuse au Gabon. Ces entretiens nous ont permis de formuler des questions analytiques :

— Q1 : Quelles sont les caractéristiques épidémiologiques de la tuberculose ? À travers cette question, nous voulons décrire le profil (proportions, variations) de la tuberculose en fonction des caractéristiques des patients tuberculeux.

— Q2 : Quelle est la répartition spatiale et temporelle des patients ? L'identification des zones géographiques les plus touchées par tuberculose est essentielle pour les autorités sanitaires. Car, elle permet de mieux orienter les actions de riposte.

3.2.2.2 Exploration des données

Afin d'apporter des réponses aux questions analytiques, à la phase d'exploration de données, nous avons réalisé un ensemble d'activités. Ces activités sont présentées ci-après :

— *Collecte des données* : Les données utilisées pour la production de la narration des données sont celles stockées dans l'entrepôt de données du SOLAP-TB proposé dans ce travail de thèse (section 3.1). Il s'agit des données sociodémographiques et cliniques qui ont été recueillies dans 7 968 dossiers médicaux de patients mis sous traitement antituberculeux à l'hôpital spécialisé de Nkembo à Libreville. Elles ont été complétées avec les données géographiques des limites administratives (quartiers et districts) de la région sanitaire Libreville-Owendo-Akanda notre zone d'étude.

— *Traitement des données* : Les données collectées ont été traitées afin de résoudre plusieurs problèmes de qualité des données (sous-section 3.1.3.3). L'outil Talend Open studio a été utilisé pour cette activité.

— *Analyse des données* : Les données ont été analysées à l'aide de l'outil de BI Tableau [4] utilisé au niveau du SOLAP-Tuberculose ainsi qu'avec le logiciel d'analyse statistique RStudio. Pour cela, nous avons réalisé de nombreuses requêtes tout en notant les résultats.

  Nous présentons ci-dessous les résultats d'analyses réalisées. La tuberculose touchant principalement les personnes vulnérables, nous avons commencé par étudier les distributions des dimensions les plus concernées par cet aspect (profession, âge, statut VIH et géographie). Les résultats obtenus ont suscité de nouvelles interrogations.

  La répartition des patients selon le statut professionnel, illustrée par la figure 43, a montré une prédominance d'élèves-étudiants (21,94%) et des sans emplois (17,81%). En outre, la proportion de patients avec un statut professionnel inconnu était de (22,39%).

  L'âge moyen des patients était de 33 ans. La répartition des cas par groupe d'âge

4. https ://www.tableau.com/

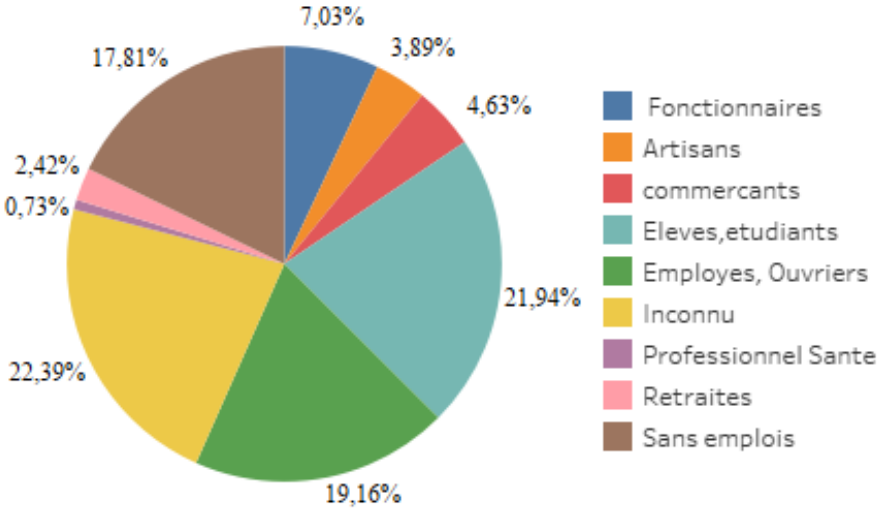


FIGURE 43 – Statut professionnel des patients tuberculeux

(voir la partie supérieure de la figure 44) a montré un plus grand nombre parmi les jeunes adultes (20-34 ans) (46%).

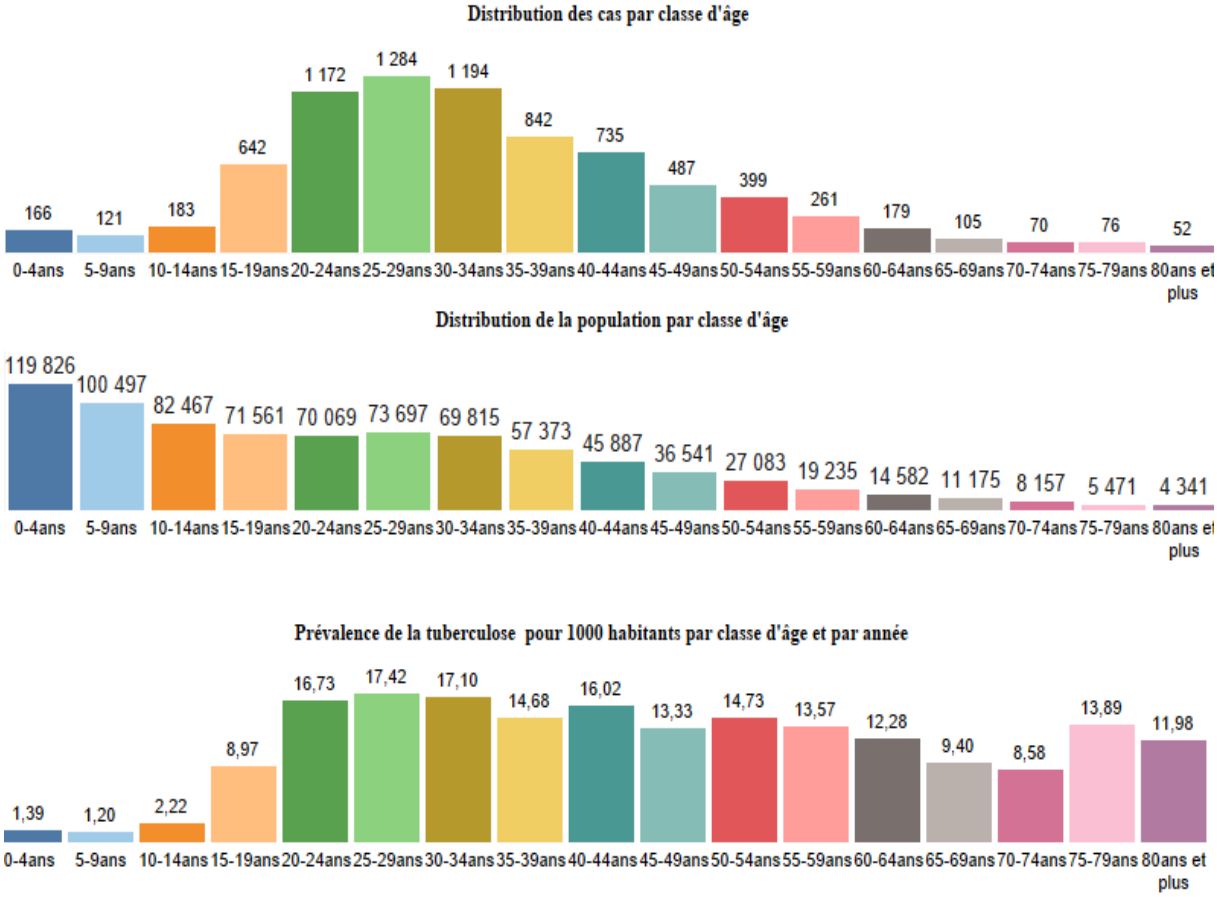


FIGURE 44 – Distribution des cas par groupe d'âge (en haut) ; pyramide des âges dans la région d'étude (au milieu) ; prévalence par groupe d'âge, sur 3 ans, pour 1000 habitants (en bas).

Au niveau géographique, les distributions spatiales des patients par quartier (Figure 45) n'ont pas montré de corrélation spatiale. Afin d'aller plus loin, nous avons étudié la distribution spatio-temporelle par typologie de quartier (Figure 46), où les tendances sont plus marquées (p-value = 0,01). D'autres sources de données ont été sélectionnées pour cette analyse, et des outils statistiques supplémentaires ont été utilisés (par exemple : Rstudio).

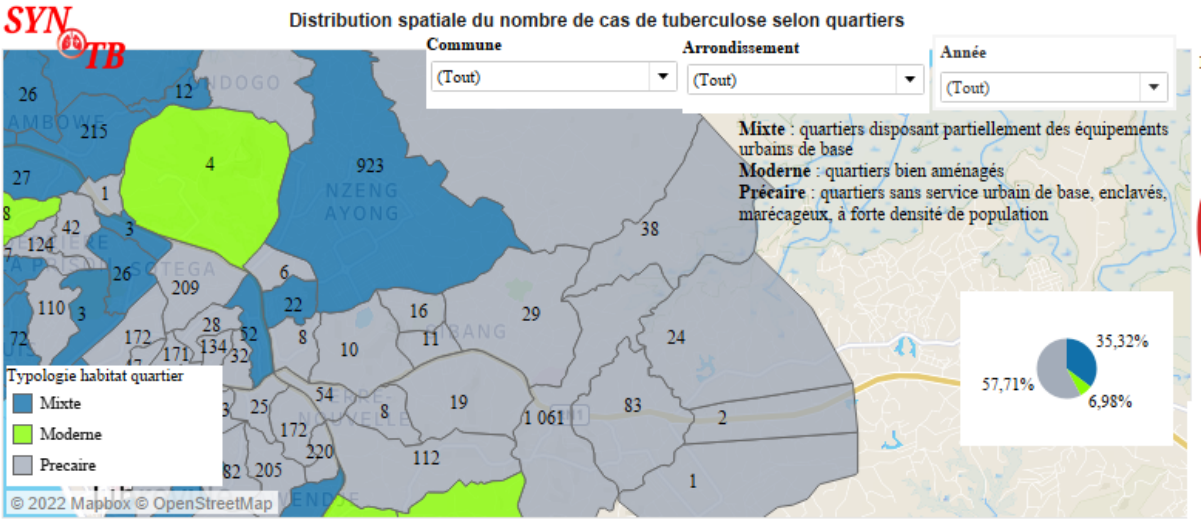


FIGURE 45 – Distribution spatiale des cas de tuberculose par quartier

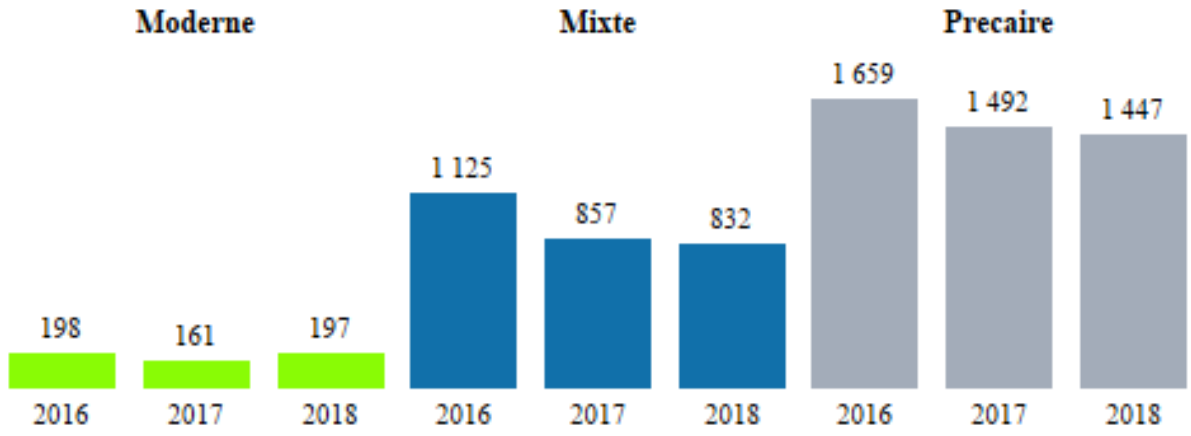


FIGURE 46 – Évolution du nombre des cas par typologie de quartier et par année

La figure 47 montre que, entre 2016 et 2018, avec 27,64% des patients, le sixième arrondissement de la commune de Libreville a été plus touché par la tuberculose (Figure 47 à gauche).

La répartition des patients selon l'issue thérapeutique a montré que 74,60% étaient perdus de vue. Aussi, l'analyse bivariée a montré que le résultat thérapeutique perdu de vue était plus associé aux sujets hommes (p-value=0,0002).

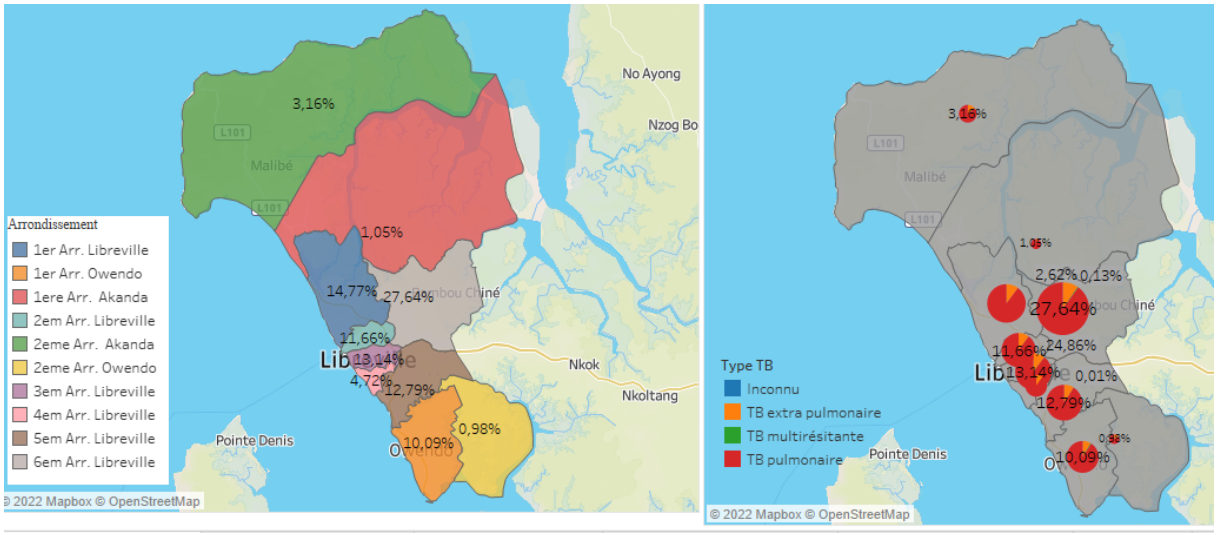


FIGURE 47 – Distribution des cas (%) de tuberculose par arrondissement

Une fois l'ensemble des données analysées, nous avons par la suite procédé à leur interprétation.

— *Interprétation des résultats* :

À la phase d'analyse, nous avons relevé certains résultats d'analyse qui sont spécifiques au Gabon, notamment les fortes proportions d'élèves-étudiants et de perdus de vue. Nous faisons focus sur l'interprétation de ces deux résultats.

Concernant la forte proportion d'élèves-étudiants parmi les patients, elle pourrait être liée à la forte consommation de drogue et alcool dans les écoles. En effet, une étude réalisée au Lycée Paul Idjendjé Gondjout de Libreville 119 a montré que parmi les élèves interrogés, âgés de 14 à 20 ans, 40% consommaient de l'alcool, 20% fumaient du tabac et 15% du cannabis. Pour mettre ces résultats en perspective, compte tenu de la jeunesse relative de la population gabonaise, nous avons étudié la pyramide des âges de la zone d'étude (partie centrale de la figure 44), tirée du recensement général de la population et de l'habitat de 2013 au Gabon (DGS, 2013), qui a montré que plus de la moitié de la population a moins de 22 ans.
Ces deux résultats nous ont permis de calculer la prévalence (nombre de patients pendant la période d'étude pour 1000 habitants) par groupe d'âge (partie inférieure de la figure 44). Nous avons relevé que les enfants étaient peu touchés par la tuberculose, mais que les jeunes adultes étaient à peine plus touchés que les adultes plus âgés. Enfin, il faut noter que le faible nombre de patients âgés ne permet pas de formuler des conclusions statistiquement valables.

Pour ce qui est de la forte proportion de patients perdus de vue, ce résultat pourrait s'expliquer par le fait des ruptures de stocks toujours fréquentes de médicaments antituberculeux qui amènent les patients à ne plus venir consulter au centre de diagnostic et de traitement (CDT). Par exemple, durant la période

d'étude (entre 2017 et 2018), le Gabon a connu une rupture de stock de médicament antituberculeux qui a durée presque un an. Aussi, l'accessibilité géographique peut également être une des causes de cette forte proportion de patients PDV. Car, par exemple, dans notre zone d'étude (Libreville-Owendo-Akanda), les patients qui sont pour la plupart issus des quartiers précaires sont souvent confrontés au problème de moyen de transport pour rallier le seul CDT (Hôpital Spécialisé de Nkembo) chargé de leur prise en charge situé parfois à une vingtaine de Km de leur lieu d'habitation. Enfin, une mauvaise éducation thérapeutique des patients ainsi que de leur entourage et la stigmatisation pourraient également expliquer cette forte proportion.

— *Vérification des résultats* : Les résultats de l'analyse ont montré que plus de la moitié des patients (57,71%) provenaient de quartiers précaires (voir figure 45). Ce résultat est en accord avec celui de (Alloghe et al., 2006).
D'autres résultats ont également été vérifiés dans la littérature. Par exemple, les hommes sont plus touchés par la tuberculose (61,60% des patients), ce qui est confirmé par d'autres études (Melki et al., 2015; Larbani et al., 2017) ; les formes pulmonaires de la tuberculose ont été observées chez une grande majorité des patients (90%) et les formes résistantes aux antituberculeux étaient marginales (0,28%), ce qui est en accord avec les connaissances métiers.
Par contre, d'autres résultats s'écartent de l'état de l'art. Par exemple, la proportion de patients perdus de vue, (75%) est beaucoup trop élevée par rapport à d'autres pays de la sous-région de l'Afrique sub-saharienne (par exemple au Mali (Sylla et al., 2017), où les perdus de vue représentaient 23,6%). Cet écart devrait faire l'objet d'une étude approfondie.
Par ailleurs, la proportion de patients sans emploi (17,81%), même si elle est importante, est très faible par rapport à celles relevées dans d'autres études (Niang et al., 2018; Tékpa et al., 2019), avec respectivement 24% et 42,70%. Cette proportion doit être complétée par le nombre de cas dont statut professionnel est inconnue (cf. Figure 43).
En s'appuyant sur la comparaison des résultats d'analyse à l'état de l'art, il faut procéder par la suite à l'évaluation du risque épidémique.
— *Évaluation du risque épidémique* : Il existe un risque de propagation de la tuberculose au sein de la population étudiante. Car, nous avons observé que cette population est plus infectée par la tuberculose. Aussi, le très fort nombre de patients tuberculeux perdus de vue peut conduire à une épidémie de tuberculose généralisée dans le pays. Enfin, les zones géographiquement défavorisées sont plus à risque tuberculose. Car, plus de la moitié des patients tuberculeux suivi à l'hôpital de Nkembo provenaient des quartiers précaires.

— *Formulation des messages* : Les résultats obtenus lors de l'analyse et de l'interprétation des données, après validation et analyse des risques, permettent de formuler un ensemble de messages à communiquer au public :
  — M1 : En dehors des patients de statut professionnel inconnu (22,39%), les étudiants sont les plus touchés par la tuberculose (21,94%). L'influence de la consommation de drogues est une piste à explorer.
  — M2 : La proportion de patients sans emploi (17,81%) est plus faible que dans les autres pays africains.
  — M3 : Une forte proportion des patients (45,81%) a entre 20 et 34 ans.
  — M4 : Les valeurs de prévalence les plus élevées (de 16,73 à 17,42 pour 1000 habitants, en 3 ans) concernent également les patients âgés de 20 à 34 ans, mais la prévalence diminue légèrement pour les patients plus âgés.
  — M5 : Les enfants sont peu touchés par la tuberculose (la prévalence est inférieure à 2,22 pour 1000 habitants, en 3 ans).
  — M6 : La proportion de patients perdus de vue (74,60%) est alarmante. Elle est plus élevée que dans les autres pays africains.
  — M7 : Il n'y a pas de corrélation spatiale ou spatio-temporelle.
  — M8 : L'évolution spatiale et temporelle est corrélée à la typologie des quartiers (p-value = 0,01), les quartiers précaires et mixtes étant significativement plus touchés, mais présentant une légère tendance à la baisse.
  — M9 : Une très grande majorité des patients (90%) souffre de la forme pulmonaire, ce qui est conforme aux connaissances métiers. Une tuberculose multirésistante a été enregistrée chez 0,28% des patients.
  — M10 : La distribution par forme clinique des perdus de vue est très similaire à celle de l'ensemble des patients.
  — M11 : Parmi les patients, il y a une prédominance masculine (61,60%).
  — M12 : La proportion de patients dont le statut VIH est inconnu est très élevée (65%). Cela ne permet pas de prendre en compte ce critère dans le profil des patients.
  — M13 : Le sixième arrondissement de Libreville enregistre la plus forte proportion de patients (27,64%).
  — M14 : Dans tous les arrondissements, la tendance générale est à une légère baisse de la prévalence. Cependant, deux arrondissements (1er d'Akanda et 2ème d'Owendo) présentent des tendances particulières et des variations plus importantes, expliquées en partie par leur faible population.
  — M15 : Deux quartiers enregistrent les plus forts taux de patients TB, Les-PK et Nzeng-Ayong, avec respectivement 13,32% et 11,58% de patients. Cependant, la prévalence pour 1000 habitants est plus élevée au quartier Alibadeng (quartier dit mixte) 34,17 cas pour 1000 habitants en 2016, 49,46 cas pour 1000 habitants

en 2017 et 38,67 cas pour 1000 habitants en 2018. Il est suivi d'un autre quartier mixte, Nzeng-Ayong avec 33,56 cas pour 1000 habitants en 2016, 21,59 cas pour 1000 habitants en 2017 et 22,68 cas pour 1000 habitants en 2018.

— M16 : Plus de la moitié des patients (57,71%) sont issus de quartiers pauvres ou précaires. Par contre, la prévalence de la TB pour 1000 habitants est plus élevée dans le quartier mixte avec 15,28 cas pour 1000 habitants.

— M17 : Dans la répartition des patients perdus de vue par type de quartiers, on retrouve la même tendance que pour l'ensemble des patients

— M18 : La tuberculose touche toutes les classes sociales et tous les sexes, la plupart des patients étant des adultes et vivant dans des quartiers pauvres ou précaires.

— M19 : Le profil des patients perdus de vue, en termes de sexe, d'âge, de profession et de localisation, est très similaire à celui des autres patients.

Parmi ces messages, ceux qui peuvent permettre de convaincre facilement les autorités sanitaires seront sélectionnés. Cette activé se réalise à la phase *structuration de la narration.*

3.2.2.3 Structuration de la narration

À cette phase, nous avons d'abord assemblé et ordonné les messages dans un cadre cohérent et logique afin de faciliter leur compréhension et d'attirer facilement l'attention du public. Ensuite, nous avons défini le public et sélectionné les messages à transmettre. Enfin, nous avons choisi la structure narrative et organisé les messages à communiquer.

— *Choix de l'audience* : Le public cible est constitué d'experts de la santé publique qui sont des décideurs.

Sélection des messages : Tous les messages formulés à la phase d'exploration des données sont sélectionnés pour être communiqués au public cible. Cette activité permet de procéder au choix des messages à même de convaincre facilement les autorités sur la situation épidémiologique de la tuberculose. L'objectif est de les amener à prendre conscience de l'ampleur de la maladie afin qu'ils prennent rapidement des décisions de riposte.

— *Choix de la structure narrative* : En s'appuyant sur les travaux de (El Outa et al., 2020), nous avons organisé la trame de l'histoire en plusieurs actes. Les actes sont composés d'épisodes, chacun racontant un message. L'intrigue est organisée en huit actes, listés dans la première colonne du tableau 3.9. Le premier acte introduit l'intention de l'étude. Le deuxième acte présente les messages saillants, et les six actes suivants se concentrent chacun sur une dimension décrivant les patients, respectivement, le statut professionnel, le sexe, l'âge, le statut VIH et la géographie. Le dernier acte présente les conclusions et les recommandations.

Nous avons opté pour cette structuration, car celle-ci facilite la compréhension

du profil d'un patient, en fonction des différentes dimensions qui le composent. On s'attend à ce que le public regarde d'abord l'introduction et la présentation générale, mais qu'il puisse ensuite naviguer entre les actes suivants en fonction de ses besoins. L'ordre de navigation n'a pas d'incidence sur les conclusions.
Cette structure est bien connue dans la narration moderne sous le nom de Martini-Glass (Segel and Heer, 2010). Elle combine plusieurs types d'interactivité de manière équilibrée : L'auteur fixe son parcours narratif (actes I et II) et le lecteur interagit et explore les chemins disponibles pour mieux comprendre les données (actes III à VIII). La navigation dans l'histoire est ainsi flexible, en fonction des besoins du public. Enfin, les messages ont été mis en correspondance avec les actes, comme le montre le tableau 3.9.

Acte	**Titre**	**Messages inclus**
I	Introduction	
II	Présentation générale	M6, M9, M10, M16
III	Profession	M1, M2
IV	Genre	M11
V	Âge	M3, M4, M5
VI	Statut VIH	M12
VII	Géographique	M7,M8,M13,M14,M15, M17
VIII	Conclusion	M18, M19

TABLE 3.9 – Structuration de la narration

3.2.2.4 Présentation

À cette phase, nous avons réalisé les activités suivantes :

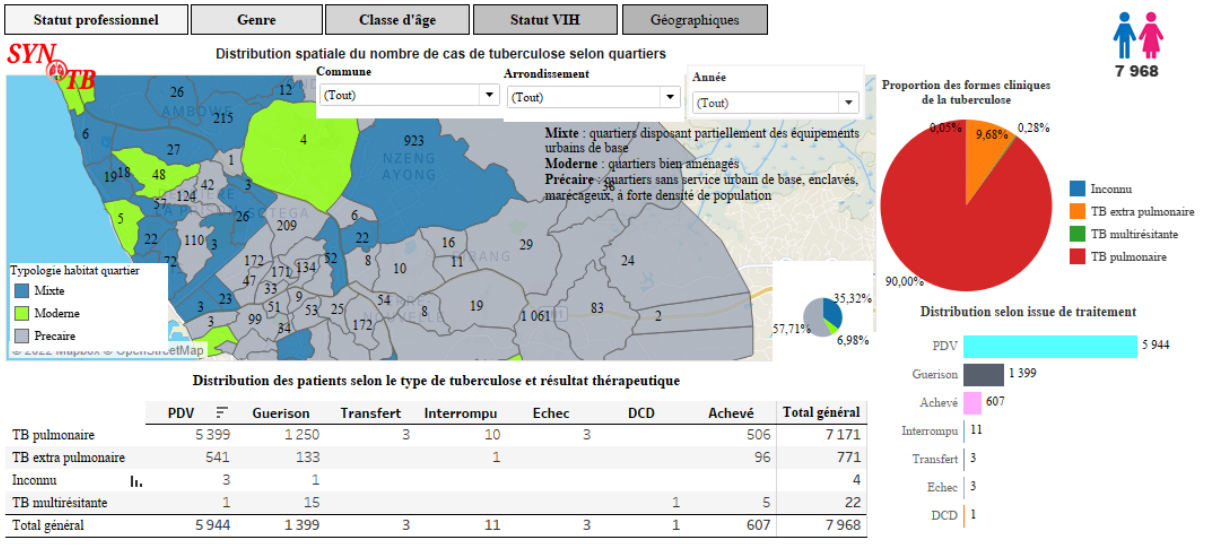
— *Choix du type de narration visuelle* : Nous avons conçu et mis en œuvre deux versions de la narration des données avec un rendu visuel différent : (i) un récit interactif, composé de tableaux de bord interactifs interconnectés (Par exemple en appliquant un filtre sur le tableau de bord qui présente la dimension âge, celui-ci peut s'appliquer sur les tableaux de bord décrivant d'autres dimensions comme le statut VIH, le type de tuberculose), et (ii) une vidéo [5], capturant une navigation particulière à travers le récit interactif, avec des explications audio. Nous avons utilisé l'outil BI Tableau pour le rendu de la narration interactive et OBS Studio pour l'enregistrement de la vidéo.

— *Élaboration des tableaux de bord* : Nous avons testé plusieurs possibilités de visualisation, mis en place les différents graphiques, cartes et tableaux, et les avons complétés par des textes explicatifs, des effets visuels et des explications audio. Les différents tableaux de bord ayant été conçus au niveau du SOLAP-TB, les différents de types de visualisations (carte, tableaux, graphiques) ainsi que les couleurs

5. https ://www.youtube.com/watch ?v=u$_K$oBWc$_q$JUt = 226s

à utiliser (par exemple : pour la carte des cas de TB, les zones en rouge foncé sont celles dans lesquelles on dénombre le plus grand nombre de patients tuberculeux) ont été discutées et validées avec les responsables du programme national de lutte contre la tuberculose au cours des séances de démonstrations.

Chaque acte est représenté par un tableau de bord interactif, à l'exception de l'acte VII (géographie) qui, devant afficher plusieurs cartes, est décomposé en plusieurs tableaux de bord. Ensuite, chaque épisode (concernant un message) est représenté par un ou plusieurs artefacts visuels (par exemple, des graphiques, des cartes, des tableaux, du texte, des images, du son). À titre d'exemple, la figure 48 correspond au tableau de bord de l'acte II. Il présente (i) la distribution géographique du nombre des patients TB par quartier (carte), (ii) la répartition du nombre de patients par type de tuberculose (diagramme en secteurs), (iii) le nombre de patients selon le résultat thérapeutique (graphique en barres) et (iv) la distribution des patients selon le type de tuberculose et les résultats thérapeutiques (tableau croisé). L'interface offre la possibilité de filtrer les indicateurs par quartier, arrondissement et année, pour une focalisation spatio-temporelle approfondie, et de zoomer sur une zone de la carte. En outre, ce tableau de bord donne accès à d'autres tableaux de bord dans lesquels d'autres visualisations (statut professionnel, genre, classe d'âge, statut VIH et géographiques) sont présentées.



	PDV	Guerison	Transfert	Interrompu	Echec	DCD	Achevé	Total général
TB pulmonaire	5 399	1 250	3	10	3		506	7 171
TB extra pulmonaire	541	133		1			96	771
Inconnu	3	1						4
TB multirésitante	1	15				1	5	22
Total général	5 944	1 399	3	11	3	1	607	7 968

FIGURE 48 – Tableau de bord acte II

3.2.3 Discussion et leçons apprises

Le processus d'élaboration est inspiré de modèles et de processus de l'état de l'art (Lee et al., 2015; Chen et al., 2020; El Outa et al., 2022). Cependant, plusieurs particularités du contexte applicatif nous ont amenés à enrichir ce processus avec les activités spécifiques au processus d'intelligence épidémique que recommande l'organisation mondiale de la santé.

Cette section présente les principaux enseignements tirés de cette adaptation.

Après l'étape d'analyse des données, il est souvent important de comparer les résultats avec l'état de l'art. Ainsi, les décideurs peuvent juger quelles dimensions du profil des patients sont en accord avec la situation dans d'autres pays, pour lesquelles des actions communes peuvent être mises en place, et lesquelles concernent la population gabonaise. De même, les résultats obtenus doivent subir des tests approfondis afin de prouver leur valeur statistique. Le public cible étant majoritairement scientifique, ces résultats peuvent être communiqués dans le récit.

Contrairement aux récits saisonniers (fréquents dans le journalisme de données), dans les récits scientifiques, les questions analytiques ne sont pas toutes connues à l'avance. Au contraire, de nouvelles questions peuvent surgir au cours de l'analyse des données. Des itérations entre les phases de définition des objectifs et d'exploration des données sont souvent nécessaires. De nouvelles trouvailles peuvent également avoir un impact sur les messages précédents et nécessiter une mise à jour.

La composante géographique est très importante pour évaluer l'étendue spatiale et spatio-temporelle des problèmes de santé. La restitution sous forme de cartes est à privilégier, mais aussi, les corrélations spatiales.

La narration de données doit permettre une navigation interactive entre les tableaux de bord. Il existe différents profils parmi les décideurs. D'une part, on retrouve des besoins variés en termes de dimensions et d'indicateurs étudiés, pour lesquels une organisation thématique est parfaitement adaptée. D'autre part, les autorités sanitaires ont besoin d'une lecture plus complète et guidée du récit. Le défi consiste à trouver un bon équilibre pour le rendu, à la fois guidé et interactif.

3.2.4 Conclusion

Le processus de production d'une narration de données dans le domaine de la santé publique nécessite de prendre en compte les particularités de ce domaine. C'est pourquoi, le processus classique de narration de données doit être enrichi à certaines phases, comme l'exploration des données, en intégrant les activités spécifiques au processus d'intelligence épidémique recommandé par l'Organisation Mondiale de la Santé, à savoir : la vérification et l'évaluation du risque épidémique.

Aussi, contrairement à d'autres types de narration (en journalisme, en politique…, et aux travaux antérieurs en santé publique) qui visent le grand public, la narration produite à travers le processus proposé vise les experts de la santé publique, qui sont des décideurs et s'appuie sur des données réelles stockées dans un entrepôt de données multidimensionnelles du SOLAP-TB proposé (section 3.1). Grâce aux résultats comparés à l'état de l'art et communiqués par des messages, les experts pourront mieux cibler les actions de lutte contre la tuberculose. Par exemple, pour réduire l'incidence des perdus

de vue tuberculeux, sensibiliser les médecins responsables de centres de diagnostic et de traitement (CDT) de la tuberculose sur la nécessité d'améliorer le suivi et la prise en charge des patients.

En plus des autorités gabonaises (responsable du programme national de lutte contre la tuberculose), la narration a été présentée à des experts de santé publique d'autres pays africains aux Journées Camerounaises d'Informatique Médicale (JCIM 2021), via une communication orale (Ondzigue Mbenga et al., 2021). Nous espérons que cette initiative servira à inspirer d'autres équipes pour reproduire l'expérience dans d'autres domaines de la santé, notamment pour faciliter la compréhension de la situation épidémiologique d'autres maladies infectieuses (par exemple le choléra, Ebola, etc.) qui sévissent encore dans nombreux pays d'Afrique.

3.3 Systèmes multi-agents de la tuberculose : SMA-TB

Contexte :

Les épidémiologistes font souvent recours au système multi-agents pour simuler et étudier la dynamique de transmission des maladies infectieuses dans la population. Concernant la tuberculose, les modèles proposés dans la littérature sont simplistes, c'est-à-dire qu'ils reposent sur des modèles classiques en épidémiologie SIR (Susceptibles-Infectés-Rétablis) et SEIR (Susceptibles-Exposés-Infectés-Rétablis) qui sont le plus souvent utilisés. Aussi, aucun des systèmes proposés n'utilise les données réelles (cliniques et géographiques) stockées dans un entrepôt de données spatiales pour générer les populations d'individus et l'environnement spatial du modèle. Enfin, ces modèles permettent de simuler la propagation inter-individus de la tuberculose et non les politiques sanitaires de lutte antituberculeuse.

Objectif :

Notre objectif est de proposer un modèle de système multi-agents de la tuberculose complet, c'est-à-dire qui tient compte de tous les états d'un individu dans le cycle de prise en charge de la maladie. Aussi, pour générer les populations d'individus et l'environnement de simulation du modèle, les données réelles stockées dans l'entrepôt de données spatiales multidimensionnelles du SOLAP-TB seront utilisées pour les simulations. Enfin, étant donné que ce modèle est destiné aux autorités sanitaires, il permettra de simuler les politiques sanitaires de lutte antituberculeuse. Cet objectif représente notre troisième verrou scientifique baptisé système multi-agents de la tuberculose (SMA-TB).

Dans cette section, nous présentons de façon détaillée le processus de conception du

système multi-agents de la tuberculose.

3.3.1 Architecture du SMA-TB proposé :

La figure 49 présente l'architecture du SMA-TB. Elle est composée d'un entrepôt de données spatiales multidimensionnelles et d'une plate-forme de simulation multi-agents.

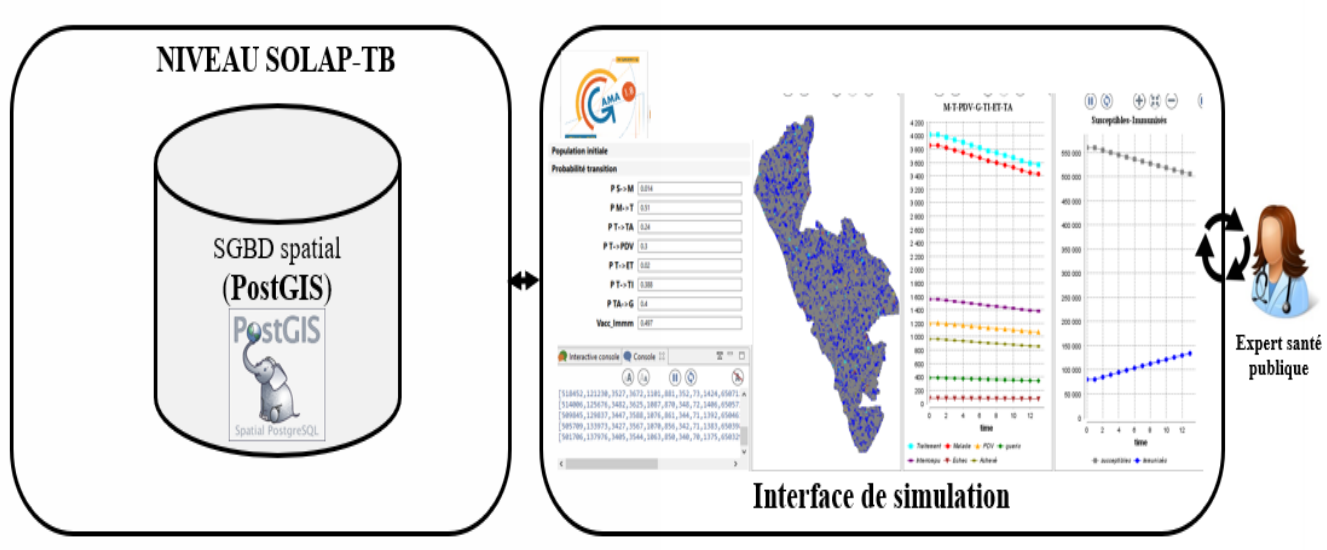


FIGURE 49 – Architecture SMA-TB

L'entrepôt de données est situé au niveau du système d'information géographique décisionnel de la tuberculose (SOLAP-TB) (Section 3.1). Ce qui permet de réaliser des simulations à partir des données réelles (données cliniques des patients) mise à jour via la plate-forme Web d'enregistrement des patients tuberculeux et stockées dans cet entrepôt. Aussi, les individus du modèle ainsi que l'environnement géographique de simulation (la région sanitaire Libreville-Owendo-Akanda) sont générés à partir de ces données.

La plate-forme multi-agents GAMA a servi à implémenter et à simuler le modèle SMA-TB proposé. Le langage GAML de GAMA qui est un langage orienté objet dérivé du JAVA a été utilisé comme langage de programmation. Dans ce langage, la définition d'un modèle s'organise en quatre blocs : 1) global, pour la déclaration des variables, les actions, la dynamique et les initialisations globales ; 2) environnement, les propriétés de l'environnement global ; 3) Entités, pour la définition des agents et enfin 4) Expérimentations, il s'agit du contexte d'exécution des simulations, en définissant par exemple leurs entrées et sorties. Plusieurs expériences peuvent être définies dans un même modèle.

L'objectif du SMA-TB est d'aider les experts en santé publique qui sont des décideurs à mieux comprendre la dynamique spatiale de la propagation inter-individus de la tuberculose dans la région sanitaires Libreville-Owendo-Akanda du Gabon.

À la sous-section 3.3.2, nous décrivons de façon détaillée les différentes phases de conception du modèle du système multi-agents de la tuberculose. Ensuite, à la sous-section 3.3.3, nous implémentons le modèle et présentons les résultats des simulations de différentes politiques sanitaires de lutte. Nous terminons par une conclusion.

3.3.2 Conception du modèle SMA-TB

L'implémentation du modèle du SMA-TB s'est fait en plusieurs étapes : la collecte des données, la construction du modèle, la codification, les tests de simulations selon différents scénarios qui sont des politiques sanitaires de lutte et l'analyse des résultats des simulations.

3.3.2.1 Collecte des données

Les données que nous avons utilisées pour la simulation sont celles stockées dans l'entrepôt de données spatiales multidimensionnelles situé au niveau SOLAP-TB. Il s'agit des données cliniques des patients et des données géographiques (quartiers qui constituent notre environnement de simulation avec leur population) de la région sanitaire Libreville-Owendo-Akanda.

3.3.2.2 Construction du modèle

Méthodologie : L'étude de l'état de l'art que nous avons réalisé sur les systèmes multi-agents de la tuberculose a montré que les modèles proposés (Prats et al., 2016; Vila Guilera, 2017; Balama and Corneille, 2017) sont moins détaillés. C'est-à-dire qu'ils ne tiennent pas compte de l'ensemble des états que peut prendre un individu dans le processus de contamination et de prise en charge de cette maladie. Aussi, ces modèles sont plus destinés à l'étude de la inter-individus de la tuberculose à des pas de temps courts (le jour).

Le modèle que nous proposons dans ce travail, modélise l'évolution des états d'un individu dans tout le processus de contamination et de prise en charge de la tuberculose comme une chaîne de Markov. Il s'agit d'un modèle qui a pour objectif de permettre aux experts de la santé publique qui sont des décideurs de prédire à plus long terme (plus de 10 ans) la situation épidémiologique future de cette maladie en simulant des politiques sanitaires à un pas de temps de six mois. Le pas de temps de simulation de six mois est lié au fait que l'objectif est de simuler les politiques sanitaires de lutte. Ce qui nécessite un pas de temps plus long par rapport aux simulations journalières (Balama and Corneille, 2017) qui sont dédiées à la compréhension de la transmission inter-individus de la tuberculose. Ce choix est également lié au fait que le traitement d'un individu infecté par la tuberculose dure six mois au minimum.

Le modèle tient également compte du renouvellement de la population par les naissances et l'immigration. Enfin, les individus décédés pour cause de tuberculose et d'autres

causes sont retirés du modèle.

Résultat : La figure 50 présente le modèle général d'états du système multi-agents de la tuberculose (SMA-TB). Dans ce modèle, nous considérons qu'un individu peut passer par neuf états :

- Susceptible (S) : individus pouvant être infectés par la TB et souffrir de la tuberculose maladie ;
- Malade (M) : individus ayant développé la tuberculose maladie ;
- Traitement (T) : individus souffrant d'une tuberculose maladie mis sous traitement antituberculeux ;
- Traitement Achevé (TA) : individus ayant achevé le traitement antituberculeux ;
- Traitement Interrompu (TI) : individus ayant interrompu le traitement antituberculeux ;
- Échec de Traitement (ET) : individus ayant fait un échec de traitement ;
- Guéris (G) : individus déclarés guéris après un traitement ;
- Perdu de Vue (PDV) : individus qui ne sont pas revenus au CDT après interruption du traitement.
- Immunisés (Imm) : individus qui acquièrent une immunité temporaire à la suite d'une vaccination au BCG (Bacille de Calmet et Guérin).

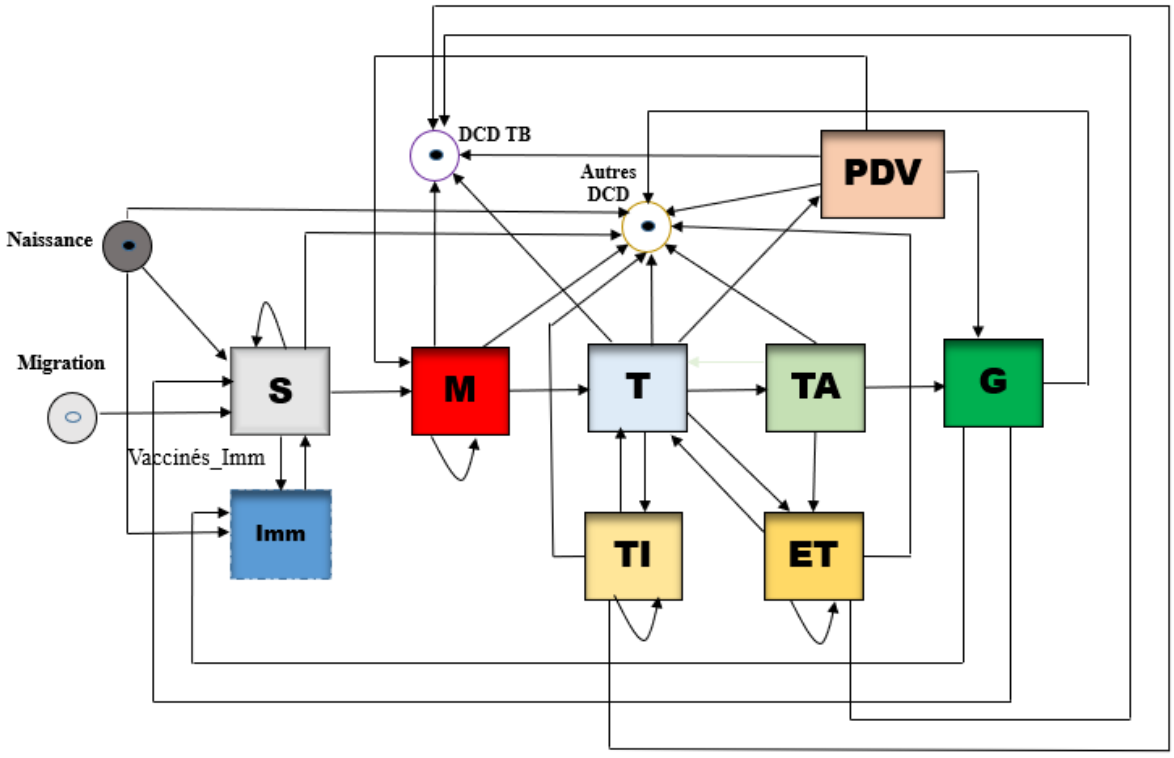


FIGURE 50 – Diagramme d'états du SMA-TB

Pour ce qui est de l'immunité des individus, nous considérons qu'un individu peut

acquérir une immunité temporaire après une guérison ou une vaccination. Dans les paragraphes qui suivent, nous décrivons de façon détaillée chaque état du modèle.

Le tableau 3.10 présente les probabilités de transition entre état du modèle, leur mode de calcul pour certains et les sources pour les autres. Il faut relever que les chiffres des probabilités de transition que nous n'avons pas pu obtenir à partir du SOLAP-TB ou dans la littérature ont été soumises aux experts de la lutte antituberculeuse qui ont trouvé ces estimations très probables.

Naissance et migration : A la naissance, les nouveaux-nés acquièrent une immunité temporaire via une vaccination au BCG. Ils peuvent également décéder pour une autre cause ou devenir par la suite susceptible d'être infecté par la tuberculose. Selon les données de l'annuaire statistiques sanitaires du Gabon 2020, le nombre de naissances vivantes enregistrés dans la région sanitaire Libreville-Owendo-Akanda était de 16 779, soit environ 8 390 par semestre.

Aussi, nous tenons également compte du fait que les populations peuvent migrer. Ces données de population permettent de renouveler la population du modèle (Figure 51). Selon une étude de l'Organisation Internationale du Travail réalisée en 2020 (Foka and Werquin, 2020), le Gabon accueille 352 600 immigrants. La province de l'Estuaire est la principale destination. Car, elle accueille 192 766 étrangers sur les 352 600 résidant au Gabon, soit 56%. Dans cette province, Libreville (notre zone d'étude appelée dans l'administration sanitaire Libreville-Owendo-Akanda) reste le principal pôle d'attraction des immigrants.

Le taux d'enfants vaccinés au BCG est de 70% (MinSanté, 2020) dans la région sanitaire Libreville-Owendo-Akanda. Aussi, selon la littérature, sur une population d'enfants vaccinés au BCG,70% acquièrent une immunité (Tolou, 2014). C'est pourquoi, la probabilité de transition de naissance à l'état immunisé (P Naissance ->Imm) est le produit du pourcentage d'enfants vaccinés par celui de vaccinés au BCG qui peuvent acquérir l'immunité. Pour la probabilité de transition de naissance à autre décès (P Naissance->DCD-Autre), nous avons considéré le taux de mortalité au Gabon qui est de 7 décès pour 1000 habitants (Banque-Mondiale, 2020). Enfin, nous avons calculé, la probabilité de transition de naissance à l'état susceptible (P Naissance->S) : 1-(P Naissance->DCD-Autre + P Naissance->Imm).

Nous passons maintenant à l'étude des états susceptible, malade, en traitement, traitement achevé, guéri, échec de traitement, traitement interrompu, perdu de vue et immunisé et aux probabilités de transition à partir de ces états.

Susceptible (S) : Il s'agit des individus qui peuvent développer une tuberculose maladie (Figure 52). Ceux-ci peuvent également développer une immunité (Imm) temporaire après vaccination au BCG ou décéder pour une autre cause (Autre DCD).

TABLE 3.10 – Probabilités de transition des états du modèle, mode de calcul ou valeur et source

Probabilité	**Mode de calcul ou valeur**	**Source**
P Naissance->DCD-Autre	Taux mortalité=0,007	(Banque-Mondiale, 2020)
P Naissance->Imm	Taux vaccinés BCG x Taux vaccinés BCG immunisé=0,7x0,70=0,49	(Tolou, 2014) et (Min-Santé, 2020)
P Naissance->S	1-(P S->DCD-Autre+PNaissance->Imm) =1-(0,007+0,49)=0,503	Calculé
P S->M	Nombre-Total-casTB (Données collectées)/Population(DGS, 2013)/nombre d'individus malades au semestre précédent soit 7968/572444 ; valeur initiale à 0,014. Cette valeur est multipliée par *Nombre de malades initialement /Nombre d'individus en traitement au semestre précédent(1491)*	Données SOLAP-TB (recalculée à chaque itération).
P eS	1-(P S->M+mortalité+P S->Imm) =1-(0,014+0,007+0,014) ; valeur initiale à 0,965	Recalculé à chaque itération
P M->DCDTB	0,045	PNLT-Gabon
P M->T	0,51	PNLT-Gabon
P eM	1-(P M->DCDTB+P M->DCD-Autre+P M->T) =1-(0,045+0,007+0,51)=0,438	Calculé
P T->TA	0,24	PNLT-Gabon
P T->ET	0,02	PNLT-Gabon
P T->PDV	0,3	PNLT-Gabon
P T->TI	1-(P M->DCDTB+P M->DCD-Autre+ P T->TA+P T->ET+P T->PDV)= 1-(0,045+0,007+0,24+0,02+0,3)=0,388	Calculé
P T->ET	0,02	PNLT-Gabon
P TA->G	0,40	PNLT-Gabon
P TA->T	1-(P M->DCD-Autre+P TA->G+P T->ET) =1–(0,007+0,02+0,4) =0,573	Calculé
P G->Imm	0,1	Estimation
P G->S	1-(P M->DCD-Autre+P G->Imm) =1-(0,007+0,1)=0,893	Calculé
P ET->T	1-(P M->DCD-Autre+P M->DCDTB+ P T->ET) =1-(0,007+0,045+0,02) =0,9283	Calculé
P TI->T	1-(P M->DCD-Autre+P M->DCDTB+P eTI) =1-(0,007+0,045+0,083)=0,865	Calculé
P PDV->G)	0,1	(Abu-Raddad et al., 2009)
P PDV->M	1-(P M->DCDTB+P M->DCD-Autre+P PDV->G)= 1-(0,045+0,007+0,1)=0,848	Calculé
P-Imm->S	0,001	Estimation

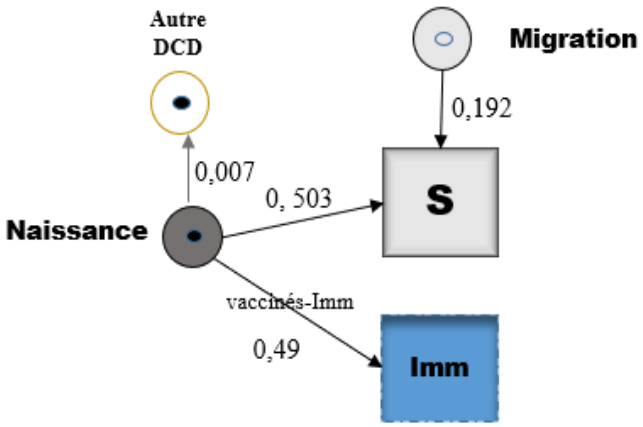


FIGURE 51 – Probabilités de transition à partir naissance

Concernant les paramètres, nous avons d'abord calculé la probabilité de transition de l'état susceptible (S) à l'état malade (M) (P S->M). Il faut relever que le nombre d'individus malades est défini au début de l'initialisation du modèle à partir du nombre de malades en traitement extrait de l'entrepôt de données multidimensionnelles du SOLAP-Tuberculose sur six mois. Mais, au cours de la simulation qui se fait à un pas de temps de six mois, la probabilité de transition de l'état susceptible à l'état malade ne doit pas être une valeur constante. En effet, elle est fonction du nombre total d'individus malades dans la population. Autrement dit, plus le nombre de malades décroît, plus cette probabilité décroît. Aussi, n'ayant pas d'information précise sur le lien entre ces deux chiffres, nous avons donc retenu une fonction linéaire qui passe par les deux points (0,0) et tient compte des chiffres extraits du SOLAP.

Nous avons calculé la probabilité de transition de l'état susceptible à l'état malade de la manière suivante : (Nombre-Total-casTB/population zone d'étude) x (Nombre de malades au semestre/ nombre de malades en traitement au semestre précédent).

Par contre, nous avons utilisé le même taux mortalité (0,007). Ensuite, nous avons calculé la probabilité de transition de l'état S à Imm : P S->Imm=Nombre d'enfants vaccinés BCG LBV-OWE-AKA/Population LBV-OWE-AKA /2 x Taux d'individus qui acquièrent l'immunité après vaccination. Dans la littérature, n'ayant pas pu obtenir des informations sur la population immunisée à la tuberculose dans notre zone d'étude, nous avons considéré la population d'immunisés au semestre. Enfin, la probabilité d'individus qui restent à l'état susceptible est déduite des autres probabilités.

État Malade (M) : Il s'agit des individus qui ont développé la tuberculose maladie. Ces individus sont par la suite mis sous traitement, c'est le passage de l'état M à l'état T (M->T) (Figure 53). Aussi, en étant malade, il peut décéder de tuberculose (DCD TB) ou d'une autre cause (Autre DCD). Cependant, certains individus malades peuvent ne

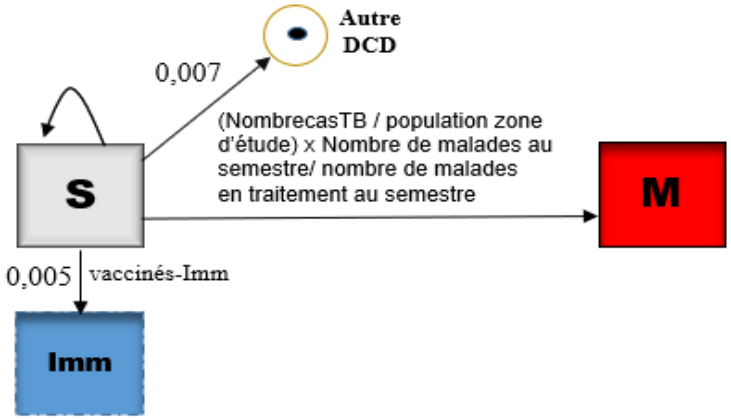


FIGURE 52 – Probabilités de transition à partir de l'état susceptible

pas être mis sous traitement antituberculeux au niveau du CDT. Car, comme le montre l'étude de (Sissoko et al., 2013), des patients tuberculeux (18,5%) disaient avoir recours aux tradipraticiens lorsqu'ils étaient atteints de tuberculose.

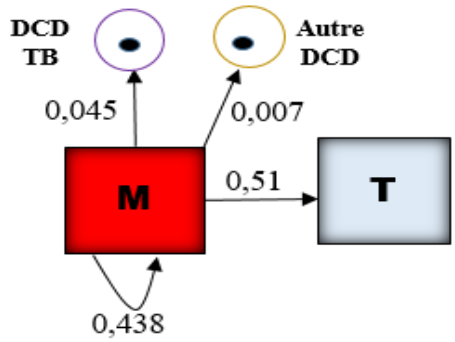


FIGURE 53 – Probabilités de transition à partir de l'état malade

Concernant les paramètres, nous considérons la même probabilité de transition de l'état malade à DCD-Autre (P M->DCD-Autre). Par contre, selon les données du programme national de lutte contre la tuberculose, la probabilité de transition de l'état M à DCDTB (P M->DCDTB) qui correspond du taux de mortalité liée à la tuberculose est de 4,5%. Celle de la transition de l'état malade (M) à l'état traitement (T) (P M->T) est de 51% (Fonds-Mondial, 2020). À partir de ces probabilités P M->DCDTB, P M->DCD-Autre et P M->T, nous avons déduit la probabilité (P-eM) d'un individu de demeurer à l'état M : P eM = 1-(P M->DCDTB+PM->DCD-Autre+P M->T).

État Traitement (T) : L'état traitement (T) regroupe les individus malades qui sont mis sous traitement (Figure 54). Parmi ces individus, certains peuvent interrompre le traitement (TI), décéder de tuberculose (DCD TB), décéder d'une autre cause (Autre DCD), faire un échec de traitement (ET), achever leur traitement (TA) ou être perdu de

vue (PDV).

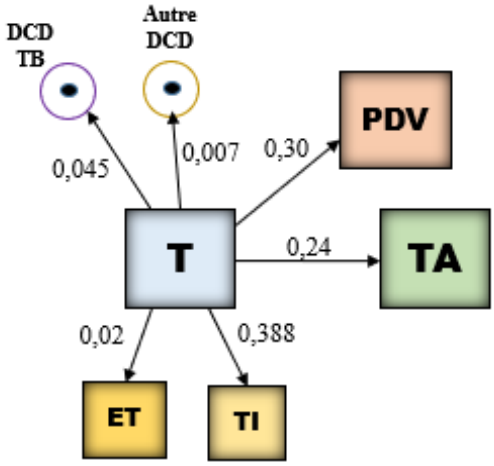


FIGURE 54 – Probabilité de transition à partir de l'état traitement

Concernant les paramètres, nous avons considéré la même probabilité de transition de malade à décédé de tuberculose (P M->DCDTB). Par contre, les probabilités de transition de traitement (T) à traitement achevé (TA) (P T->TA), T à échec de traitement (ET) (P T->ET) et T à perdu de vue (PDV) (P T->PDV) sont respectivement de 24%, 2% et 30% (PNLT-Gabon). À partir de ces différentes probabilités de transition, nous avons déduit la probabilité de transition de l'état T à l'état traitement interrompu (TI) (P T->TI) : 1-(P M->DCDTB+P M->DCD-Autre+P T->TA+ P T->ET). Il faut relever que le traitement ayant une durée six mois, un individu ne peut pas demeurer à l'état T.

État traitement achevé (TA) : Il s'agit des individus ayant achevé un traitement antituberculeux d'une durée de six mois (Figure 55). Une fois le traitement achevé, les individus peuvent être déclarés guéris (G) cliniquement par un médecin. Aussi, un échec de traitement (ET) peut être constaté et l'individu peut également mourir à cause d'une autre affection (Autre DCD). Enfin, le reste des individus sont reversés à l'état de traitement (T). Cette transition peut être liée au fait que le premier traitement administré au patient n'a pas permis la guérison. Ce qui conduit à administrer à celui-ci un nouveau traitement qui peut être de première ligne ou de deuxième ligne selon les cas.

Concernant les paramètres, nous avons considéré les mêmes probabilités pour les décès pour autre cause (0,007) et les échecs de traitement (0,02). Par contre, la probabilité de transition de l'état TA à l'état G (P TA->G) est de 40% (PNLT-Gabon). À partir de ces probabilités, nous avons déduit la probabilité de transition de l'état TA à T (P TA->T) : TA->T=1-(P M->DCD-Autre+P TA->G+P T->ET). Il s'agit des individus ayant achevé le traitement et qui ne sont pas déclarés guéris ou ayant fait un échec de traitement. Ces individus passent à nouveau à l'état T, parce qu'ils vont bénéficier d'un

nouveau traitement qui est différent du premier.

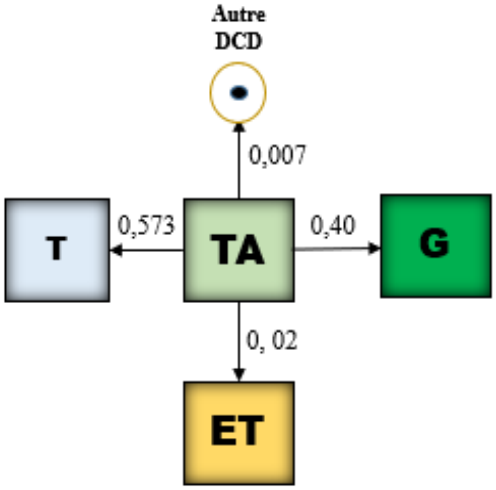


FIGURE 55 – Probabilité de transition à partir de l'état traitement achevé

État Guéri (G) : Il s'agit des individus qui ont suivi un traitement antituberculeux d'une durée de 6 mois et qui sont déclarés cliniquement guéris (G) (Figure 56). Ces individus peuvent mourir d'une autre affection (Autre DCD), redevenir susceptible (S) ou acquérir une immunité relative (Imm).

Pour les paramètres, nous avons considéré la même probabilité de décès pour autre cause P M->DCD-Autre (0,007). Par contre, ne disposant pas de plus d'informations sur les guéris immunisés et en tenant compte du fait que la vaccination contre la tuberculose ne confère pas une immunité complète, nous avons retenu une probabilité de 10% d'individus guéris qui seront considérés comme immunisés. Ensuite, nous avons déduit la probabilité de transition de l'état G à l'état S (P G->S) : 1-(P M->DCD-Autre+P G->Imm). Il s'agit d'individus guéris qui redeviennent susceptibles d'être infectés par la tuberculose.

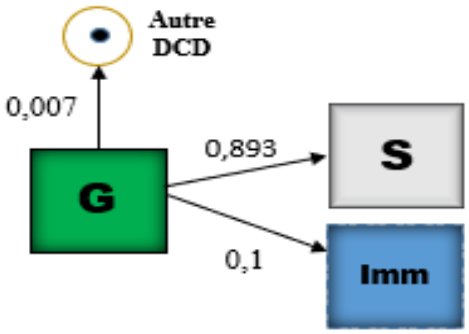


FIGURE 56 – Probabilités de transition à partir de l'état Guéris

État Échec de Traitement (ET) : Il s'agit des individus dont le traitement antituberculeux n'a pas donné lieu à une guérison (Figure 57). Dans la très grande majorité des cas, ces individus sont remis en traitement (T). Aussi, ils peuvent également décéder de tuberculose (DCD TB) ou d'une autre affection (Autre DCD). Enfin, certains individus restent à l'état d'échec de traitement (ET).

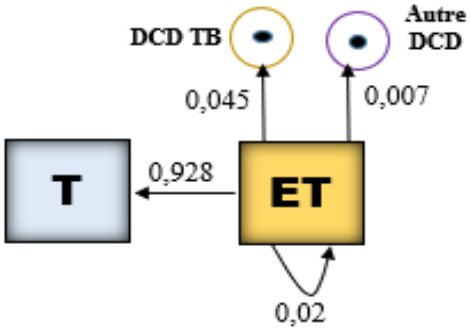


FIGURE 57 – Probabilités de transition à partir de l'état échec de traitement

En ce qui concerne les paramètres, nous avons considéré les mêmes probabilités d'échec, de traitement (0,02), de DCD TB (0,045) et DCD-Autre (0,007). Ensuite, nous avons déduit la probabilité de transition de l'état ET à l'état T (P ET->T) : 1-(P M->DCD-Autre +P M->DCDTB+P T->ET).

État Traitement Interrompu (TI) : Il s'agit des individus ayant interrompu le traitement antituberculeux (Figure 58). Pendant cette interruption, ils peuvent décéder de tuberculose ou d'une autre affection. Par contre, lorsqu'un individu interrompt son traitement, il devrait le recommencer (T).

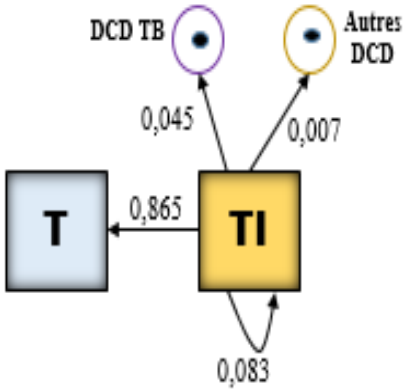


FIGURE 58 – Probabilités de transition à partir de l'état traitement interrompu

Concernant les paramètres, la probabilité de demeurer à l'état traitement interrompus (P eTI) est à 0,083, qui correspond au taux de traitement interrompu retrouvé dans (Kombila et al., 2017). Aussi, nous avons considéré les mêmes probabilités pour les décès TB (0,045) et les décès autre cause (0,007). À partir de ces probabilités, nous avons déduit la probabilité d'individus ayant interrompu le traitement et qui sont reversés à l'état traitement (T), P TI->T : 1-(P M->DCD-Autre+ P M->DCD TB+P eTI).

État Perdu de Vue (PDV) : Les individus Perdus de Vue (PDV) sont reversés à l'état malade (M). Il s'agit d'individus malades qui ont abandonné définitivement leur traitement. Nous estimons que ces individus peuvent guérir de façon spontanée (P PDV->G), mourir de tuberculose (P PDV->DCD TB) ou d'une autre affection (P PDV->DCD-Autres).

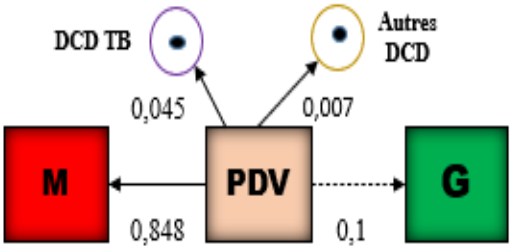


FIGURE 59 – Probabilités de transition à partir de l'état perdu de vue

Concernant les paramètres, nous avons considéré les mêmes probabilités pour DCD TB (0,045) et pour DCD-Autres (0,007). Par contre, nous avons considéré qu'un PDV peut être guéris de façon spontanée (P PDV->G) à une proportion de 0,1 (Abu-Raddad et al., 2009). Enfin, nous avons déduit la probabilité de transition de l'état PDV à l'état malade (M) P PDV->M : 1-(P M->DCDTB+P M->DCD-autre+P PDV->G).

État Immunisé : Les individus qui ont été vaccinés au BCG peuvent perdre leur immunité (Figure 60) et redevenir susceptible d'être infectés par la tuberculose. Aussi, n'ayant pas plus d'informations sur la proportion de ces individus (P Imm->S), nous avons considéré une proportion de 0,1%.

Nous venons de décrire de façon détaillée les différents états de modèle SMA-TB. La section 3.3.3 présente le processus d'implémentation de ce modèle ainsi que les résultats

3.3.3 Implémentation du modèle

Méthodologie : Nous avons implémenté le modèle avec la plateforme open source de simulation multi-agents GAMA. Le langage GAML a été utilisé pour la codification.

Le tableau 3.11 présente comment nous avons estimé les populations d'initialisation du modèle pour chaque état. Nos données ayant été recueillies en milieu hospitalier, nous

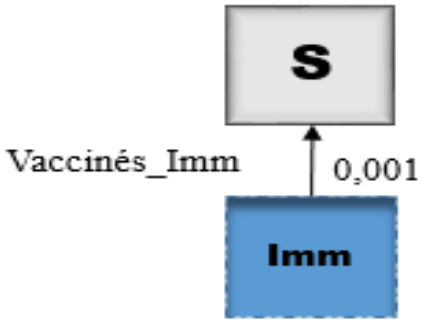


FIGURE 60 – Probabilité de transition de l'état immunisé à l'état Susceptible

avons estimé les populations d'individus malades, traitements achevés, échecs traitements, traitements interrompus et perdus de vue en fonction du nombre d'individus en traitement à l'aide des probabilités de transition de la chaîne de Markov.

La figure 61 présente l'écran du simulateur SMA-TB. Il est décomposé en cinq blocs :

1. Les populations et valeurs initiales des transitions entre état d'un individu : Le tableau 3.11 présente la population par état lors de l'initialisation du modèle et le tableau 3.12 les valeurs initiales des transitions entre état. Ces valeurs peuvent être modifiées par l'expert en santé publique par rapport à la politique sanitaire qu'il souhaite simuler. Seule la valeur de probabilité de transition de S->M ne peut être modifiée. Car, elle est calculée en fonction du nombre d'individus malades au cours de la simulation.
2. Le résultat de la simulation selon le nombre d'individus S, Imm, M, T, G, PDV, TI, ET et TA. Ce résultat est récupéré automatiquement dans un fichier au format CSV et peut donc être manipulé avec un tableur comme Excel ou un logiciel d'analyse statistique.
3. Une interface de visualisation des résultats des interactions entre les différents individus sur le SIG environnement de simulation du modèle. Par exemple, à l'initialisation du simulateur, nous pouvons visualiser les individus susceptibles (en gris), malades (en rouge), perdus de vue (en jaune), en traitement (en cyan) localisés dans les quartiers. Ces données sont chargées à partir de l'entrepôt de données spatiales multidimensionnelles situé au niveau du SOLAP-TB.
4. Un graphique qui permet de suivre l'évolution du nombre d'individus selon leur état (S, M, T, G, PDV, TI, ET et TA) au cours de la simulation (graphique à gauche).
5. Un graphique qui permet de suivre l'évolution du nombre d'individus S et Imm au cours de la simulation (graphique à droite).

Nous avons d'abord réalisé la première simulation avec les paramètres initiaux du modèle (scénario 1). Ensuite, nous avons simulé différentes politiques sanitaires possibles

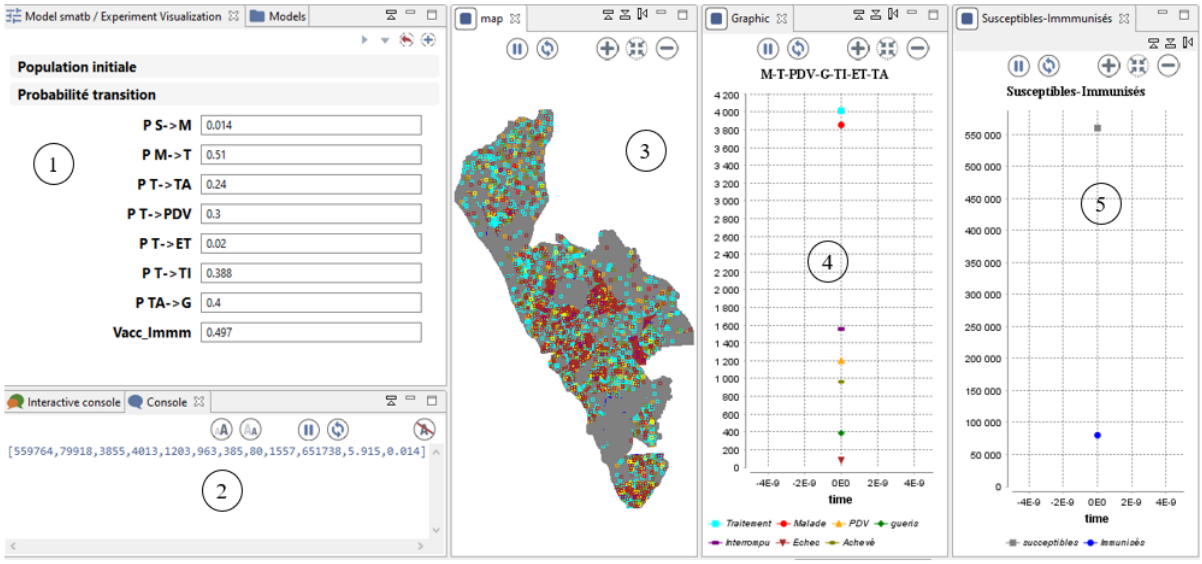


FIGURE 61 – Écran lors de l'initialisation du simulateur avec les données issues du SOLAP en 2016

TABLE 3.11 – Mode d'estimation de la population initiale du modèle par état

État	**Mode d'estimation**	**Population**
S	Pop zone étude x taux-S (échantillon de 70%)	559764
M	T x (1-P M->T)/P M->T	3855
T	Nombre de patients traités x P M->T	4013
TA	T x P T->TA	963
G	TA x P TA->G	385
ET	T x P T->ET	80
TI	T x P T->TI	1557
PDV	T x P T->PDV	1203
Imm	Population x P S->Imm	79918
Pop zone étude		817777

(scénarios 2, 3 et 4) à un pas de temps de six mois. Nous avons retenu le pas de temps six mois non seulement parce que le traitement de la tuberculose dure six mois (cas de la tuberculose sensible) mais aussi, parce que le modèle proposé a pour objectif de simuler les politiques sanitaires de lutte antituberculeuse.

Ci-après, nous présentons les résultats des simulations obtenus selon les quatre scénarios retenus pour ce travail : (i) utilisation des paramètres initiaux, (ii) diminution de la probabilité de transition à l'état PDV de 30% à 10%, (iii) augmentation des probabilités de transition en traitement de 51% à 70% et (iv) vaccinés immunisés de 70% à 80%.

TABLE 3.12 – Probabilités de transition initiales entre état du modèle

Probabilité transition	**Valeur**
P S->M	0,014
P M->T	0,51
P T->PDV	0,30
P TA->G	0,40
P T->TA	0,24
P T->ET	0,02
P T->TI	0,388
Taux_Vacc_Imm	0,497

Résultats :

Scénario 1, utilisation des paramètres initiaux du modèle : Nous avons d'abord réalisé une première simulation avec les paramètres initiaux du modèle. L'objectif visé était de vérifier que le code informatique du modèle reproduisait correctement la réalité du phénomène étudié, la tuberculose.

Le résultat (figure 62) présente le modèle à l'état initial et après simulation à T=40. Le résultat a montré que, lorsque le nombre d'individus malades M (en rouge) diminue, il y a également une diminution du nombre d'individus en traitement T (en cyan), perdus de vue PDV (en jaune), guéris G (en vert), échec de traitement ET (en violet), traitement achevé TA (vert olive) et traitement interrompu TI (en marron). Selon les spécialistes de la lutte antituberculeuse, ce résultat montré que, si le nombre de cas de tuberculose diminue dans la population, il y aura de moins en moins d'individus tuberculeux en traitement. Ce qui entraînera une diminution des issus thérapeutiques PDV, traitement achevé, guérison et échec de traitement.

Le tableau C.1 qui montre l'évolution du nombre d'individus M, T, PDV, TA, G, ET et TI du trimestre 1 au trimestre 20 (Soit 10 ans) se trouve en annexe C. Dans cette période de simulation, le nombre d'individus a diminué pour : M (de N=3855 à N=3068), T (de N=4013 à N=3194), PDV (de N=1203 à N=958), TA (de N=963 à N=766), G (de N=385 à N=306), ET (de N=80 à N=63) et TI (de N=1557 à N=1239). Nous relevons également que dans cette même période, la probabilité de transition de l'état S à l'état M baisse très faiblement, passant de 0,014 à 0,013. Aussi, il y a une variation de la population totale. En effet, celle-ci évolue en fonction du nombre d'individus S, M, T, PDV, G, TI, ET, TA et Imm.

Concernant les individus susceptibles (en gris) et immunisés (en bleu), le résultat de la simulation (figure 62) a montré que lorsque le nombre d'individus immunisé augmente, il y a une diminution du nombre d'individus susceptibles d'être infectés par la tuberculose.

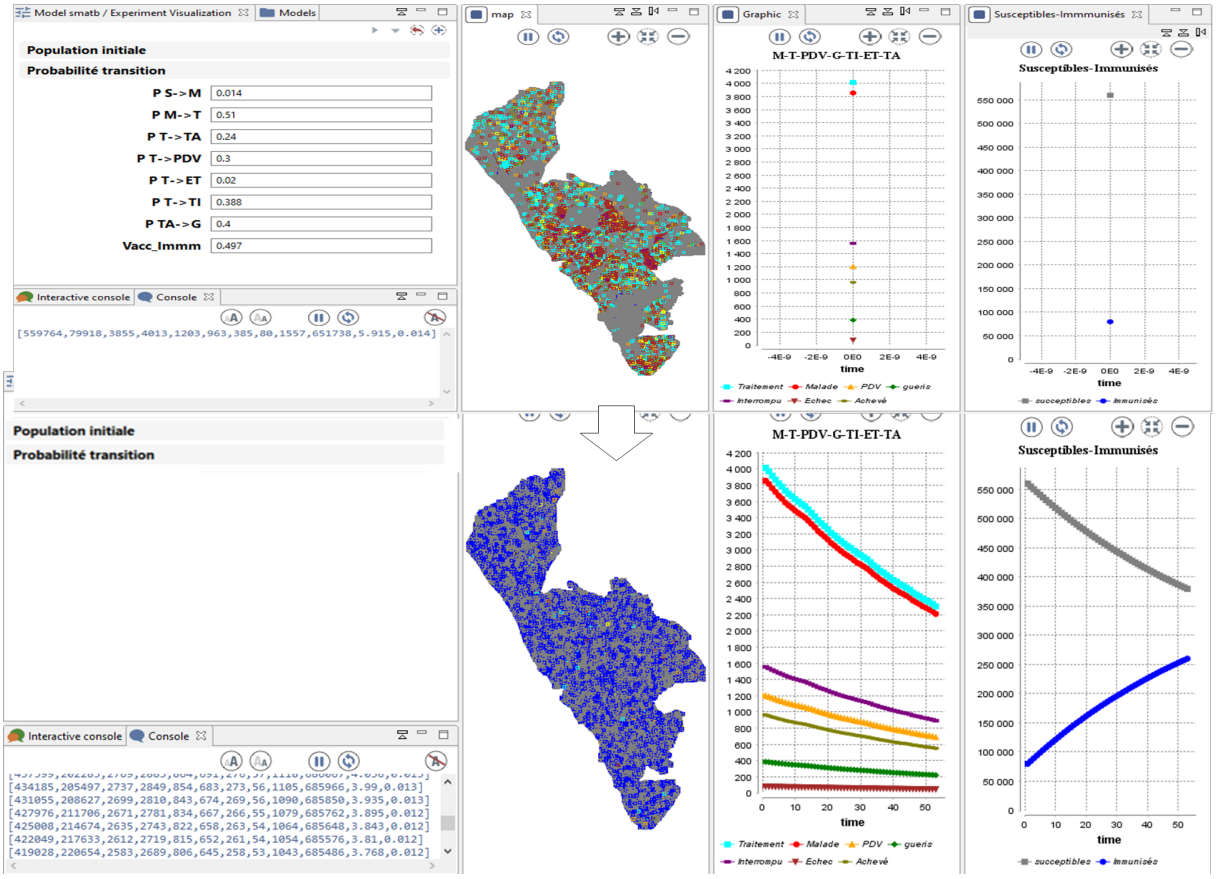


FIGURE 62 – Scénario 1 : simulation du modèle (en haut à l'initialisation du modèle et en bas le modèle après simulation à T=52)

S'agissant de la prévalence, nous avons observé une tendance baissière de celle-ci. Elle passe de 5,91 cas à 4,48 cas pour 1000 habitants sur 20 semestres (Figure 63). Ce résultat a montré que la tuberculose demeure encore endémique au Gabon. Par conséquent, si les politiques sanitaires efficaces de riposte ne sont pas mises en œuvre, ce problème majeur de santé publique ne sera pas éradiqué dans les prochaines années.

Après cette première simulation, nous avons ensuite simulé quelques stratégies de lutte antituberculeuse, à savoir : (i) la diminution de la probabilité de transition à l'état PDV de 30% à 10% (scénario 2) (ii) l'augmentation des probabilités de transition aux états traitement de 51% à 70% (scénario 3) et (iii) vaccinés immunisés de 70% à 80% (scénario 4).

Scénario 2, diminution de la probabilité de transition à l'état PDV de 30% à 10% : La figure 64 montre que sur 20 semestres de simulation, la diminution de la probabilité de transition à l'état perdu de vue de 30% à 10% entraîne une diminution du nombre d'individus malades (de N=3069 à N=2927). Mais, cette diminution reste très faible. Par contre, le nombre d'individus ayant achevé le traitement (TA) et guéris (G) augmentent. Respectivement de N=766 à N=1340 et de N=306 à N=536 (tableau C.4,

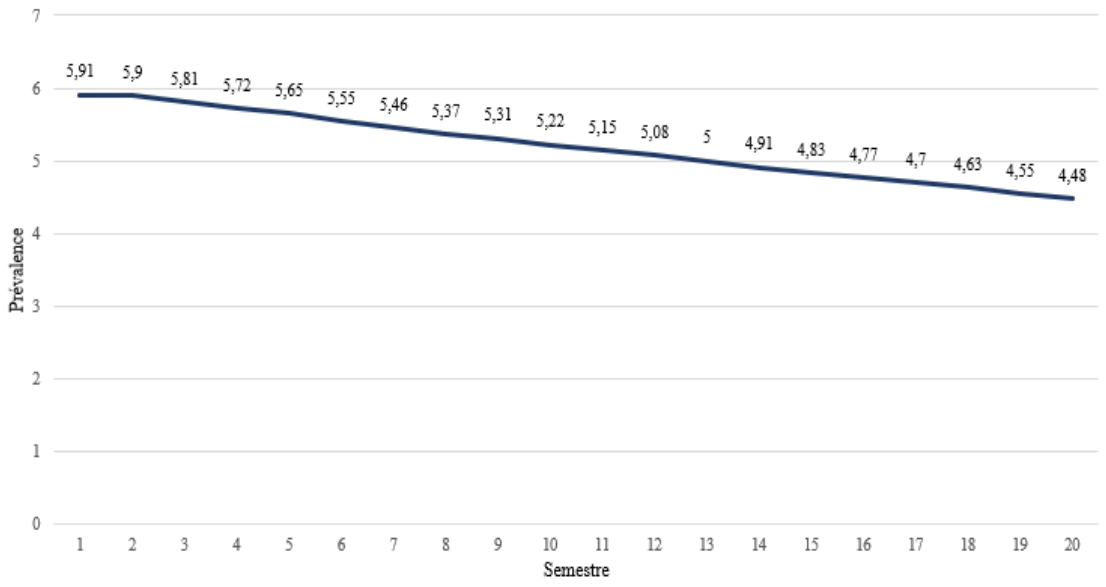


FIGURE 63 – Évolution de la prévalence pour 1000 habitants pour le scénario 1

Annexe C).

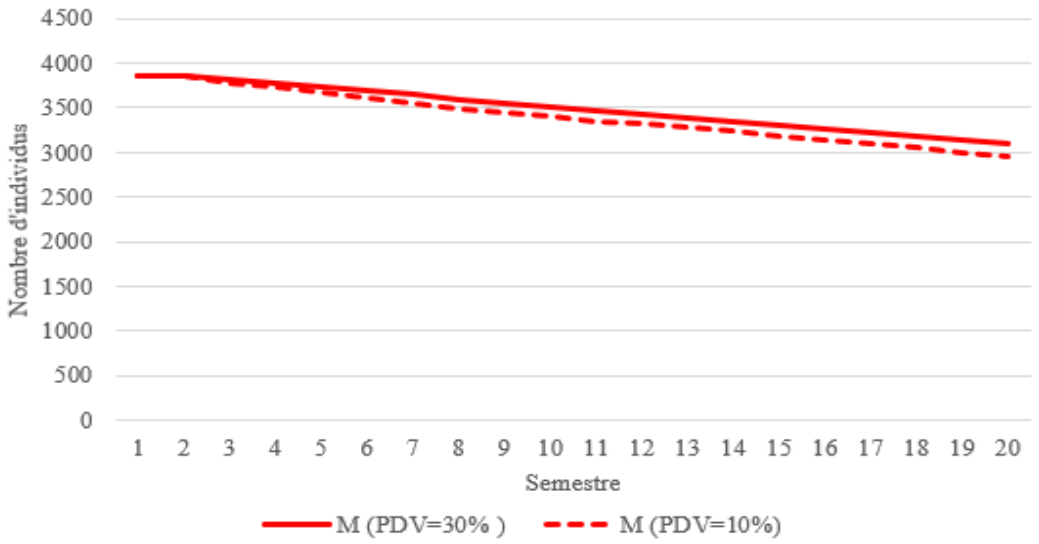


FIGURE 64 – Évolution du nombre d'individus malades (PDV=30% et PDV=10%) pour le scénario 2

Nous avons relevé que même si le nombre d'individus malades diminue très faiblement, les résultats de cette simulation sont confortés par une conclusion de l'étude de (Amadou et al., 2019). Selon les auteurs, l'éradication de la tuberculose passera en autre par une diminution du taux des perdus de vue. Cette diminution nécessite en autre une amélioration de la prise en charge des patients dans les centres de diagnostic et de traitement (par exemple : disponibilité des antituberculeux) et une sensibilisation des patients mis sous traitement antituberculeux sur les risques encourus en cas d'abandon définitif du

traitement.

Scénario 3, augmentation de la probabilité de transition à l'état traitement de 51% à 70% : Le tableau C.2 (Annexe C) montre que, lorsque la probabilité de transition à l'état traitement augmente de 51% à 70%, il y a une baisse du nombre de malades et des perdus de vue. En effet, sur 20 semestres, ces nombres ont presque diminué de moitié. Ils passaient de N=3021 à N=1557 pour les malades et de N=965 à N=439 pour les perdus de vue. Par contre, sur cette même période, il y avait une augmentation du nombre d'individus guéris de N=308 à N=773. Pour diminuer davantage le nombre d'individus perdus de vue, il faudra par exemple une meilleure éducation thérapeutique des patients qui favorise leur parfaite adhérence au traitement (Rakotomanana et al., 2000; Amadou et al., 2019). Mais aussi, un approvisionnement régulier des centres de diagnostic en antituberculeux (absence de rupture de stocks).

Scénario 4, augmentation de la probabilité de transition à l'état vacciné immunisés de 70% à 80% La figure 65 montre que sur 20 semestres de simulation, lorsque la probabilité de transition à l'état immunisé vacciné passait de 70% à 80%, le nombre d'individus immunisés vaccinés augmentait légèrement (de N=165472 à N=171465) et le nombre d'individus susceptibles diminuait aussi légèrement (de N=474210 à N=468217) (tableau C.3, Annexe C). Aussi, le nombre de malades passait de 3 194 à 2606.

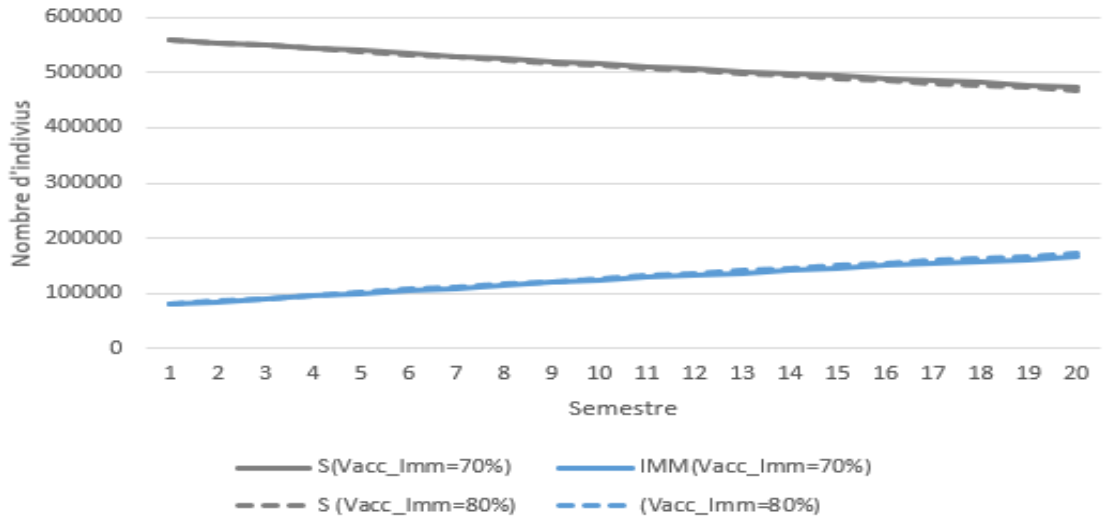


FIGURE 65 – Évolution du nombre d'individus susceptibles et vaccinés immunisés lorsque la probabilité de transition à l'état vacciné immunisé (Vacc_Imm) augmente de 70% à 80%

3.3.4 Conclusion

Dans cette section, nous avons présenté de façon détaillée le processus conception et de mise œuvre du simulateur multi-agents de la tuberculose (SMA-TB). Le modèle du SMA-TB tient compte de tous les états d'un individu dans le cycle de transmission et de prise en charge de la maladie. Il a été implémenté et simulé sous la plate-forme multi-agents GAMA. Les données descriptives et géographiques utilisées pour initialiser les simulations sont des données réelles extraites de l'entrepôt de données spatiales multidimensionnelles de la tuberculose que nous avons créé au niveau du SOLAP Tuberculose. Les probabilités de transition entre états dans la chaîne de Markov ont été définies comme suit : (i) en tenant compte de la littérature (45,45%), (ii) calculés (45,45%) et (iii) une estimation non validée (9,09%). Pour ces dernières, des études et analyses complémentaires sont nécessaires.

Nous avons simulé quelques politiques sanitaires de lutte antituberculeuse. Les résultats obtenus pour les quatre scénarios montrent qu'une seule mesure ne suffira pas à éradiquer la tuberculose. Il faudra combiner l'ensemble des mesures pour espérer mettre fin à cette maladie au Gabon.

En perspective, pour enrichir le modèle, nous envisageons d'intégrer l'état statut VIH des individus dans le cadre de la co-infection TB-VIH. Aussi, au niveau géographique, dans l'actuel modèle, la probabilité de transition de l'état susceptible à malade est identique quel que soit le quartier. Or cette probabilité devrait tenir compte du nombre de cas de tuberculose dans le quartier et les déplacements des individus entre quartier. Pour améliorer notre modèle, il faut donc un meilleur recensement des populations des quartiers et une estimation des déplacements des individus entre quartier.

Chapitre 4

Conclusion générale

4.1 Résumé des contributions

La surveillance épidémiologique est une activité de santé publique reconnue comme fondamentale pour la prévention et la lutte contre les maladies. Car, elle permet aux experts de la santé publique qui sont des décideurs d'identifier précocement de nouveaux foyers épidémiques, de surveiller l'évolution de l'épidémie, d'organiser une prise en charge appropriée des malades et d'évaluer les politiques sanitaires de lutte. Mais, pour réaliser efficacement cette activité, ils doivent disposer d'outils adaptés et utiliser de nouvelles techniques pour communiquer efficacement et aider à la prise de décision. Dans ce contexte, l'objectif de cette thèse était de concevoir un système d'information décisionnel qui puisse permettre aux experts de la santé publique de : (i) suivre et surveiller en quasi-temps réel l'évolution des indicateurs spatio-temporels de la tuberculose, (ii) améliorer la communication sur les problèmes de santé publique et (iii) simuler les politiques sanitaires de lutte antituberculeuse. Nous présentons dans les sections ci-après, un résumé de chacune de ces contributions, leurs limites et les perspectives.

4.1.1 Le SOLAP de la tuberculose

Dans la littérature, un SOLAP de la tuberculose a été proposé (Benabbou et al., 2014). Mais, le modèle conceptuel de ce SOLAP ne permet pas de réaliser des analyses selon les dimensions sociodémographiques (âge, genre, statut professionnel) et cliniques (résultat thérapeutique, type de tuberculose et statut VIH) du patient tuberculeux. Dans ce travail, le modèle conceptuel de l'entrepôt de données spatiales multidimensionnelles du SOLAP proposé pour le suivi et la surveillance de la tuberculose (SOLAP-TB) est un modèle quasi complet. Car, il intègre l'ensemble des dimensions sociodémographiques et cliniques (âge, genre, statut professionnel, statut VIH, type de tuberculose, type patient, résultat thérapeutique, temps, géographie) qui caractérisent un patient en traitement et permettent de décrire en détail le profil épidémiologique de la tuberculose.

Concernant l'architecture du SOLAP-TB, elle est composée d'une plateforme Web qui permet aux centres de diagnostic et de traitement (CDT) d'enregistrer les personnes infectées par la tuberculose. Cette plateforme a pour objectif d'amélioration la qualité dont la complétude des données utilisées dans le système de surveillance de la tubercu-

lose. L'ensemble des données sociodémographiques (âge, genre, statut professionnel, lieu résidence, CDT), temporelles et cliniques (type tuberculose, type patient, résultat thérapeutique, statut VIH) enregistrées via cette plateforme sont stockées dans l'entrepôt de données spatiales multidimensionnelles que nous avons créé. Aussi, le SOLAP-TB intègre un client d'analyse multidimensionnelle qui permet de générer des tableaux de bord dynamiques et interactifs d'indicateurs de la tuberculose (cartes, graphiques, tableaux croisés) à partir des données réelles des patients tuberculeux stockées dans l'entrepôt de données spatiales multidimensionnelles. Ces tableaux de bord que nous avons conçu avec le logiciel du Business Intelligence Tableau permettent aux autorités sanitaires de suivre et de surveiller en quasi-temps réel la dynamique spatio-temporelle de l'évolution des indicateurs de la tuberculose à différentes granularités. Les résultats d'analyses visualisés à travers ces tableaux de bord confirment des connaissances métiers (par exemple, personnes plus infectées par la forme clinique pulmonaire et quartiers précaires plus à risques tuberculose). Mais, nous avons aussi constaté quelques faits spécifiques au Gabon comme la surreprésentation de personnes infectées jeunes-adultes et élèves-étudiants.

Nous relevons que le SOLAP-TB proposé peut être facilement réutilisé par un autre pays dans le cadre de la surveillance épidémiologique de la tuberculose. Mais, cela nécessitera une intégration des données sociodémographiques et cliniques des personnes infectées par la tuberculose du pays concerné dans l'entrepôt de données spatiales multidimensionnelles.

Enfin, une extension du modèle conceptuel en intégrant d'autres maladies (par exemple, le VIH, le paludisme) peut être possible. Dans ce cas, les faits étudiés auront certaines dimensions en commun (par exemple sexe, âge, temps, géographie, statut professionnel et lieu résidence) dans un modèle en constellation d'étoiles.

4.1.2 Le processus de narration de donnés en intelligence épidémique

Nous avons constaté l'absence dans la littérature de processus de narration dédié à l'intelligence épidémique. Mais, il y a d'une part les processus généraux de narration de données (Lee et al., 2015; Chen et al., 2018; El Outa et al., 2022) et d'autre part les processus d'intelligence épidémique (WHO, 2014; Kaiser et al., 2006; Noah, 2006; Thacker et al., 2012; Savel et al., 2012; Gilbert and Cliffe, 2016; Hartley et al., 2013; Eilstein et al., 2012; Astagneau and Ancelle, 2011) qui sont proposés.

Dans ce travail de thèse, nous avons conçu un processus de narration de données en intelligence épidémique (P-N-D-I-E) qui s'inspire des processus généraux de narration de données et enrichi avec les phases particulières (vérification et évaluation du risque épidémique) issues du processus d'Intelligence Épidémique recommandé par l'Organisation Mondiale de la Santé (WHO, 2014) pour la surveillance des maladies. Ce P-N-D-I-E est

composé de quatre phases dans lesquelles un ensemble d'activités sont réalisées : (i) Objectif de la narration de données (définition de l'objectif et formulation des questions analytiques), (ii) exploration des données (collecte des données, prétraitement des données, analyse des données, interprétation des résultats, vérification des résultats, évaluation du risque épidémique et formulation des messages), (iii) structuration de la narration (choix de l'audience, sélection des messages et choix de la structure narrative) et (iv) présentation (choix du type de narration visuel et élaboration des tableaux de bord). Il faut relever que dans ce processus destiné aux experts de la santé publique, il peut y avoir des aller-retours entre phases et activités selon les besoins du producteur de la narration. Aussi, afin de mieux évaluer le risque épidémique, lors de l'activité de vérification, l'ensemble des résultats interprétés doivent être confrontés à l'état de l'art.

Le P-N-D-I-E peut être utilisé pour la production de narration de données sur n'importe quel problème de santé publique. Dans le cadre de cette thèse, en s'appuyant sur ce processus, nous avons produit une narration de données sur la tuberculose au Gabon (Ondzigue Mbenga et al., 2022).

4.1.3 Le SMA de la tuberculose

Nous avons proposé un modèle de simulation à base d'agents de la tuberculose (SMA-TB) complexe. En effet, contrairement à d'autres modèles proposés dans la littérature (Prats et al., 2016; Vila Guilera, 2017; Balama and Corneille, 2017) qui reposent sur les modèles classiques en épidémiologie SIR (Sains-Infectés- Guéris) ou SEIR (Sains-Exposés-Infectés-Rétablis), le modèle du SMA-TB est basé sur une chaîne de Markov qui tient compte de tous les états des individus pendant la transmission et le traitement de la tuberculose (susceptible, immunisé, malade, perdu de vue, guéri, en traitement, traitement interrompu, échec de traitement). Aussi, par couplage avec le système SOLAP de la tuberculose (Section 3.1), les données cliniques et géographiques (environnement de simulation) utilisées pour générer les individus et l'environnement de simulation du modèle sont des données réelles extraites de l'entrepôt de données spatiales multidimensionnelles. Ces simulations se font à des pas de temps très long (le semestre) et couvrent un espace géographique très grand, la région sanitaire Libreville-Owendo-Akanda du Gabon. Enfin, le modèle permet de simuler les politiques sanitaires de lutte antituberculeuse (par exemple, réduction de la probabilité de transition à l'état perdu de vue, augmentation des probabilités de transition aux états vaccinés immunisés et traitement). Les résultats des simulations ont montré que pour circonscrire la tuberculose au Gabon, jouer sur une seule politique de riposte ne suffit pas, mais il faudrait absolument s'attaquer à l'ensemble des problèmes de front.

4.2 Limites et Perspectives

4.2.1 Limites

Au-delà de sa contribution théorique et pratique, cette thèse renferme quelques limites qui méritent d'être relevées.

Le SOLAP-TB ne permet pas actuellement aux autorités sanitaires de surveiller les indicateurs de la tuberculose de l'ensemble des dix régions sanitaires du Gabon. Aussi, le modèle conceptuel ne tient pas compte de la dimension traitement administré aux patients. Or, cette dimension, lorsqu'elle est croisée avec une dimension comme le résultat thérapeutique peut permettre de comprendre la cause des échecs de traitement. Ce qui n'était pas l'objectif du présent travail.

Au niveau du SMA-TB, le modèle proposé ne tient pas compte des caractéristiques socio-démographiques (sexe, âge et statut professionnel) et clinique (statut VIH) des individus ainsi que de la typologie de quartier (précaire, mixte et moderne). Ces dimensions sont aussi importantes pour mieux comprendre la transmission inter-individus de la tuberculose et sa dynamique de propagation dans un espace géographique. Aussi, le modèle n'est pas centré-individu, c'est pourquoi, les simulations ont été réalisées au niveau meso (les quartiers) et la chaîne de Markov est la même pour tous les individus.

4.2.2 Perspectives

Le travail accompli durant cette thèse ouvre plusieurs perspectives de travaux futurs. Tout d'abord, pour le SOLAP-TB, à court terme, nous allons intégrer les données socio-démographiques et cliniques des personnes souffrant de la tuberculose des neufs autres régions sanitaires du Gabon. Ensuite, nous allons prendre en compte la dimension traitement du patient dans le modèle conceptuel. À long terme, nous allons envisager une extension du modèle conceptuel en intégrant d'autres « faits » qui vont permettent de mesurer d'autres maladies (par exemple le VIH, le paludisme, la Lèpre, les cancers). Enfin, une identification unique du patient tuberculeux est une piste de solution pour limiter les doubles comptes dans le nombre de personnes réellement infectées par la tuberculose au Gabon.

Pour le SMA-TB, le modèle pourra être enrichi en tenant compte des caractéristiques sociodémographiques (sexe, âge et statut professionnel, population par quartier), clinique (statut VIH) des individus (modèle centré-individus). Car, les statuts socio-démographiques et VIH sont également importants pour comprendre la dynamique de propagation inter-individus de la tuberculose. Aussi, étant donné que le modèle proposé ne tient pas compte des déplacements des individus, dans une nouvelle approche, il peut donc être intéressant de tenir compte de ces déplacements pour mieux comprendre la

dynamique de contamination inter-individu de la tuberculose. Cela nécessitera la prise en compte des lieux fréquentés par ces individus (par exemple, marché, école, lieux de culte).

Enfin, le P-N-D-I-E pourra être utilisé pour produire des narrations de données sur d'autres maladies (par exemple le VIH, diabète, les cancers du sein et du col). À court terme, nous allons produire une narration de données qui s'appuiera sur ce processus et les résultats des simulations de politiques sanitaires réalisées avec le SMA-TB.

Ces travaux de recherche ont ouvert une voie prometteuse en ce qui concerne le suivi et la surveillance de maladies par des systèmes informatiques (SOLAP, SMA). Ils doivent être poursuivis pour les améliorer en intégrant des données plus fiables et plus nombreuses, puis les généraliser à d'autres maladies et d'autres régions ou pays.

Bibliographie

Abrami, G., Amalric, M., Amblard, F., Anselme, B., Banos, A., Beck, E., Becu, N., Blanpain, B., Caillault, S., Corson, N., Daudé, E., Debolini, M. M., Delay, E., Duraffour, F., Gaudieux, A., Gaudou, B., Houet, T., Langlois, P., Laperrière, V., Lemoy, R., Louail, T., Marilleau, N., Miahle, F., Monteil, C., Moreno, D., Pivano, C., Reulier, R., Rey-Coyrehourcq, S., Rousseaux, F., Salze, P., Schmitt, C. S., Sheeren, D. D., Taillandier, P., Thierry, H., and Vannier, C. (2014). Modelisation multi-Agents appliquee aux Phenomènes Spatialises. Lecture.

Abu-Raddad, L. J., Sabatelli, L., Achterberg, J. T., Sugimoto, J. D., Longini, I. M., Dye, C., and Halloran, M. E. (2009). Epidemiological benefits of more-effective tuberculosis vaccines, drugs, and diagnostics. *Proceedings of the National Academy of Sciences*, 106(33) :13980–13985.

Akoka, J., Berti-Équille, L., Boucelma, O., Bouzeghoub, M., Comyn-Wattiau, I., Cosquer, M., Goasdoué, V., Kedad, Z., Nugier, S., Peralta, V., et al. (2008). Évaluation de la qualité des systèmes multisources une approche par les patterns. *Qualité des Données et des Connaissances*.

Alloghe, E. E., Mve, T., Ramarojoana, S., Iba Ba, J., and Nkoghe, D. (2006). Epidemiologie de tuberculose infantile au centre antituberculeux de libreville de 1997–2001. *Med trop*, 66 :469–471.

Amadou, M. L. H., Abdoulaye, O., Amadou, O., Biraïma, A., Kadri, S., Amoussa, A. A. K., Lawan, I. M., Tari, L., Daou, M., Brah, S., et al. (2019). Profil épidémiologique, clinique et évolutif des patients tuberculeux au centre hospitalier régional (chr) de maradi, république du niger. *The Pan African Medical Journal*, 33.

Amina, Z. F., Djamila, H., and Karim, B. (2011). Couplage solap et datawarehouse : Outil interactif d'aide à décision spatio-temporelle.

Amoakoh-Coleman, M., Kayode, G. A., Brown-Davies, C., Agyepong, I. A., Grobbee, D. E., Klipstein-Grobusch, K., and Ansah, E. K. (2015). Completeness and accuracy of data transfer of routine maternal health services data in the greater accra region. *BMC research notes*, 8(1) :1–9.

Andrienko, G. L. and Andrienko, N. V. (1999). Interactive maps for visual data exploration. *Int. J. Geogr. Inf. Sci.*, 13(4) :355–374.

Ariaux, B. (2000). Les sig utilisés en agriculture de précision. In *Colloque Agriculture de Précision, Educagri, Dijon*, pages 55–65.

Astagneau, P. and Ancelle, T. (2011). *Surveillance épidémiologique : Principes, méthodes et applications en santé publique*. Lavoisier.

Badariotti, D., Banos, A., and Laperrière, V. (2007). Vers une approche individu-centrée pour modéliser et simuler l'expression spatiale d'une maladie transmissible : la peste à madagascar. *Cybergeo : European Journal of Geography*.

Balama, G. and Corneille, K. V. (2017). Multi-agents modeling and simulation of the spread of tuberculosis in the city of ngaoundéré (cameroon). *International Journal of Computer (IJC)*, 27(1) :39–54.

Ban, T. Q., Duong, P. L., Son, N. H., and Van Dinh, T. (2020). Covid-19 disease simulation using gama platform. In *2020 international conference on computational intelligence (ICCI)*, pages 246–251. IEEE.

Banque-Mondiale (2020). Taux de mortalité, brut (pour 1 000 personnes) - gabon.

Barrau, D., Barthélémy, N., Kedad, Z., Laboisse, B., Nugier, S., and Thion, V. (2016). Gestion de la qualité des données ouvertes liées - État des lieux et perspectives. *Revue des Nouvelles Technologies de l'Information*.

Barreto, L. P. T. (2019). Real time data intake and data warehouse integration.

Batini, C., Cappiello, C., Francalanci, C., and Maurino, A. (2009). Methodologies for data quality assessment and improvement. *ACM computing surveys (CSUR)*, 41(3) :1–52.

Bédard, Y. (2004). Amélioration des capacités décisionnelles des sig par l'ajout d'un module solap (spatial on-line analytical processing). *Université Aix-Marseille, École Polytechnique Universitaire de Marseille, Filière Génie Industriel et Informatique*, 8.

Bédard, Y., Proulx, M.-J., and Rivest, S. (2005). Enrichissement du olap pour l'analyse géographique : exemples de réalisation et différentes possibilités technologiques. In *EDA*, pages 1–20.

Benabbou, A., Bouamrane, K., and Hamdadou, D. (2014). Mise en place d'un système d'information décisionnel spatio-temporel pour la surveillance épidémiologique.

Bernier, E. (2002). *Utilisation de la représentation multiple comme support à la génération de vues de bases de données géospatiales dans un contexte SOLAP*. PhD thesis, Université Laval.

Bernier, E. and Bédard, Y. (2006). Développement de technologies géospatiales.

Berti-Équille, L. (2004). La qualité des données comme condition à la qualité des connaissances : un état de l'art. *Revue des Nouvelles Technologies de l'Information*.

Berti-Équille, L. (2006). Qualité des données. *Techniques de l'ingénieur. Informatique HB4 (H3700)*.

Berti-Equille, L. (2007). Measuring and modelling data quality for quality-awareness in data mining. In *Quality measures in data mining*, pages 101–126. Springer.

Berti-Equille, L. (2012). *La qualité et la gouvernance des données Au service de la performance des entreprises*. Lavoisier.

Beyeme-Ondoua, J.-P. (2007). Évaluation de la qualité des données chaînées nationales du cancer colorectal. *Santé publique*, 19(6) :471–480.

Bill, R., Nash, E., and Grenzdörffer, G. (2011). Gis in agriculture. In *Springer handbook of geographic information*, pages 461–476. Springer.

Bimonte, S. (2007). Intégration de l'information géographique dans les entrepôts de données et l'analyse en ligne.

Bimonte, S. (2019). Focus - Un système OLAP pour l'analyse de données de lutte intégrée : application à la culture de l'olivier. *Sciences Eaux Territoires*, (29) :14–17.

Bimonte, S., Boulil, K., and Pinet, F. (2016a). Cohérence logique dans les systèmes olap spatiaux-un état de l'art. *Revue Internationale de Géomatique*, 26(1) :97–131.

Bimonte, S., Boulil, K., Pradel, M., André, G., and Chanet, J.-P. (2014). Analyse des indicateurs énergétiques des entreprises agricoles. une approche spatial olap. *Rev. Int. Géomatique*, 24(1) :37–65.

Bimonte, S., Hassan, A., Beaune, P., and Irstea, T. (2016b). Une architecture orientée services pour l'olap spatial. In *EDA*, pages 17–24.

Bimonte, S., Tchounikine, A., Miquel, M., and Laurini, R. (2007). Vers l'intégration de l'analyse spatiale et multidimensionnelle. In *Colloque International de Géomatique et d'Analyse Spatiale (SAGEO 2007)*.

Bommel, P. (2009). *Définition d'un cadre méthodologique pour la conception de modèles multi-agents adaptée à la gestion des ressources renouvelables*. Theses, Université Montpellier II - Sciences et Techniques du Languedoc.

Bonin, M., Augusseau, X., and Arnaud, M. (2001). Sig et statistiques pour l'analyse des dynamiques d'occupation et d'utilisation du sol : application à une commune du

parc naturel régional des monts d'ardèche et à une zone d'accueil de migration dans le sud-ouest du burkina faso.

Botsis, T., Fairman, J. E., Moran, M. B., and Anagnostou, V. (2020). Visual storytelling enhances knowledge dissemination in biomedical science. *Journal of biomedical informatics*, 107.

Bouba, F. (2015). *Système d'information décisionnel sur les interactions environnement-santé : cas de la Fièvre de la Vallée du Rift au Ferlo (Sénégal)*. PhD thesis, Université Pierre et Marie Curie-Paris VI ; Université Cheikh Anta Diop (Dakar).

Boulanger, P. and Bréchet, T. (2003). Modélisation et aide à la décision pour un développement durable : état de l'art et perspectives. *Rapport final au SPP Politiques Scientifique (AS/F5/01). Institut pour un développement Durable, Bruxelles*.

Bousquet, F., Le Page, C., and Müller, J.-P. (2002). Modélisation et simulation multi-agent. *deuxiemes assises du GDRI3*, page 26.

Brahim, M. I. (2019). L'épidémiologie mathématique.

Bui, T.-M.-A. et al. (2016). *Séparation des préoccupations en épidémiologie*. PhD thesis, Paris 6.

Câmara, G., Monteiro, A. M., Fucks, S. D., and Carvalho, M. S. (2004). Spatial analysis and gis : a primer. *Image Processing Division, National Institute for Space Research (INPE), Brasil*.

Caron, P.-Y. (1998). Étude du potentiel de olap pour supporter l'analyse spatio-temporelle.

Carpendale, S., Diakopoulos, N., Riche, N. H., and Hurter, C. (2016). Data-driven storytelling (dagstuhl seminar 16061). *Dagstuhl Reports*, 6(2) :1–27.

Chaker, W., Proulx, M., Moulin, B., and Bédard, Y. (2009). Modélisation, simulation et analyse d'environnements urbains peuplés. *Revue internationale de Géomatique*.

Chang, K.-T. (2016). Geographic information system. *International Encyclopedia of Geography : People, the Earth, Environment and Technology : People, the Earth, Environment and Technology*, pages 1–9.

Charef, A. B. (2019). Systèmes d'information décisionnels.

Chatelain, J.-L., Guillier, B., Souris, M., DUPÉRIER, E., and YEPES, H. (1995). Sig et évaluation des risques naturels : Application aux risques sismiques de quito. *Mappemonde*, 3(1995) :17–22.

Che, D. and Desenclos, J. (2002). Detection systems for infectious diseases in france. *MEDECINE ET MALADIES INFECTIEUSES*, 32(12) :704–716.

Chen, S., Li, J., Andrienko, G., Andrienko, N., Wang, Y., Nguyen, P. H., and Turkay, C. (2018). Supporting story synthesis : Bridging the gap between visual analytics and storytelling. *IEEE transactions on visualization and computer graphics*, 26(7) :2499–2516.

Chen, S., Li, J., Andrienko, G. L., Andrienko, N. V., Wang, Y., Nguyen, P. H., and Turkay, C. (2020). Supporting story synthesis : Bridging the gap between visual analytics and storytelling. *IEEE Trans. Vis. Comput. Graph.*, 26(7) :2499–2516.

Cisse, P. A., Dembele, J. M., Lo, M., and Cambier, C. (2017). Multi-agent systems for epidemiology : Example of an agent-based simulation platform for schistosomiasis. In *Agents and Multi-Agent Systems for Health Care*, pages 131–153. Springer.

Cisse, P. A. (2016). *Simulation à base d'agents de la propagation de la Schistosomiase : une approche de composition et de déploiement de modèles.* PhD thesis, Université Pierre et Marie Curie-Paris VI.

Crozat, S. (2016). Data warehouse et outils décisionnels.

Cueva, M., Kuhnley, R., Revels, L., Schoenberg, N. E., and Dignan, M. (2015). Digital storytelling : a tool for health promotion and cancer awareness in rural alaskan communities. *International journal of circumpolar health*, 74(1).

Demange, J. (2012). *Un modèle d'environnement pour la simulation multiniveau-Application à la simulation de foules.* PhD thesis, Université de Technologie de Belfort-Montbeliard.

Denis, A. (2021). Systèmes d'information géographique sig. présentation générale des sig & exemples d'applications sig pour la gestion de l'environnement et de l'agriculture.

Denis, F. and Perronne, C. (2004). Mycobacterium tuberculosis et mycobactéries atypiques. *Médi-bio.*

DGS (2013). Recensement général de la population et des logements (rgpl) de 2013 du gabon.

Dombret, M.-C. (2004). Tuberculose pulmonaire de l'adulte. *EMC-médecine*, 1(5) :406–416.

Drogoul, A. (1993). *De la simulation multi-agents a la resolution collective de problemes : une etude de l'emergence de structures d'organisation dans les systemes multi-agents.* PhD thesis, Paris 6.

Dubé, É. (2008). Conception et développement d'un service web de constitution de mini cubes solap pour clients mobiles.

Ehrlinger, L. and Wöß, W. (2022). A survey of data quality measurement and monitoring tools. *Frontiers in Big Data*, page 28.

Eilstein, D., Salines, G., and Desenclos, J.-C. (2012). Veille sanitaire : outils, fonctions, processus. *Revue d'épidémiologie et de Santé Publique*, 60(5) :401–411.

El Outa, F., Francia, M., Marcel, P., Peralta, V., and Vassiliadis, P. (2020). Towards a conceptual model for data narratives. In Dobbie, G., Frank, U., Kappel, G., Liddle, S. W., and Mayr, H. C., editors, *Conceptual Modeling - 39th International Conference, ER 2020, Vienna, Austria, November 3-6, 2020, Proceedings*, volume 12400 of *Lecture Notes in Computer Science*, pages 261–270. Springer.

El Outa, F., Marcel, P., and Peralta, V. (2021). Un modèle conceptuel de narration de données. In *17th French Days on Business Intelligence and Big Data (EDA'2021)*.

El Outa, F., Marcel, P., Peralta, V., da Silva, R., Chagnoux, M., and Vassiliadis, P. (2022). Data narrative crafting via a comprehensive and well-founded process. In *European Conference on Advances in Databases and Information Systems*, pages 347–360. Springer.

Ezegbe, B., Eseadi, C., Ede, M. O., Igbo, J. N., Aneke, A., Mezieobi, D., Ugwu, G. C., Ugwoezuonu, A. U., Elizabeth, E., Ede, K. R., et al. (2018). Efficacy of rational emotive digital storytelling intervention on knowledge and risk perception of hiv/aids among schoolchildren in nigeria. *Medicine*, 97(47).

Ferber, J. (1995). Les systèmes multi-agents : vers une intelligence collective. *InterEditions, Paris*, 322.

Ferber, J. (1997). *Les systèmes multi-agents : vers une intelligence collective*. InterEditions.

Foka, S. and Werquin, P. (2020). Potentiel de partenariats pour les compétences et la migration au gabon. *Organisation Internationale du Travail*.

Fonds-Mondial (2020). Rapport 2020 sur les résultats vih, paludisme et tuberculose. `https://www.theglobalfund.org/media/10162/corporate_2020resultsreport_report_fr.pdf`.

Fruytier, C. (2009). L
aide á la validation des données.

Gilbert, R. and Cliffe, S. J. (2016). Public health surveillance. In *Public Health Intelligence*, pages 91–110. Springer.

Goulet, M. (2012). Hiérarchiser les dimensions de la qualité des données : analyse comparative entre la littérature et les praticiens en technologies de l'information. *Essai présenté au CeFTI, Faculté des Sciences, Université de Sherbrooke, Longueuil, Québec, Canada*.

Gourmelon, F. (2003). *La contribution des SIG à la connaissance et à la gestion de l'environnement littoral*. PhD thesis, Université de Bretagne occidentale-Brest.

Gowthami, K. and Kumar, M. P. (2017). Study on business intelligence tools for enterprise dashboard development. *International Research Journal of Engineering and Technology*, 4(4) :2987–2992.

Grimaud (1998). Introduction à la modélisation et à la simulation. `https://www.emse.fr/~grimaud/Simulation/CoursRapide/CoursRapide.htm`. Accessed : 2022-07-19.

Grouls, A. (2013). Agents et systemes multi-agents : vers une synthese de ces concepts/memoire presente comme exigence partielle de la maitrise en informatique par alexandre grouls ;[directeur de recherche, roger nkambou].

Hartley, D. M., Nelson, N. P., Arthur, R., Barboza, P., Collier, N., Lightfoot, N., Linge, J., van der Goot, E., Mawudeku, A., Madoff, L., et al. (2013). An overview of internet biosurveillance. *Clinical Microbiology and Infection*, 19(11) :1006–1013.

Hermellin, E. (2016). *Modélisation et implémentation de simulations multi-agents sur architectures massivement parallèles*. PhD thesis, Université Montpellier.

Holmes, J. H., Ratshaa, B., and Steenhoff, A. P. (2013). Analyse et gestion des données. *Guide de la recherche sur le cancer en Afrique*, page 141.

IEEE (1998). Ieee standard for a software quality metrics methodology. *IEEE Std 1061-1998*, pages i–.

Inmon, W. H. (1995). What is a data warehouse. *Prism Tech Topic*, 1(1) :1–5.

Izewski, N. (2022). A netlogo covid-19 virus simulation model for determining better strategies at handling a virus outbreak.

Jabri, H., Lakhdar, N., El Khattabi, W., and Afif, H. (2016). Les moyens diagnostiques de la tuberculose. *Revue de Pneumologie Clinique*, 72(5) :320–325.

Jebb, A. T., Parrigon, S., and Woo, S. E. (2017). Exploratory data analysis as a foundation of inductive research. *Human Resource Management Review*, 27(2) :265–276.

Jennings, N. R., Sycara, K., and Wooldridge, M. (1998). A roadmap of agent research and development. *Autonomous agents and multi-agent systems*, 1(1) :7–38.

Joliveau, T. (1996). Gérer l'environnement avec des sig mais qu'est-ce qu'un sig ?/managing environment with gis but what is a gis ? *Géocarrefour*, 71(2) :101–110.

Kabachi, N. (1999). *Modélisation et Apprentissage de la Prise de Décision dans les Organisations Productives : Approche Multi-Agents*. Theses, Ecole Nationale Supérieure des Mines de Saint-Etienne ; Université Jean Monnet - Saint-Etienne.

Kabunga, S. K. (2020). *Towards hybrid stochastic modeling and simulation of complex systems in multi-scale environments with case studies on the spread of tuberculosis in Democratic Republic of the Congo*. PhD thesis.

Kaiser, R., Coulombier, D., Baldari, M., Morgan, D., and Paquet, C. (2006). What is epidemic intelligence, and how is it being improved in europe ? *Weekly releases (1997–2007)*, 11(5) :2892.

Kamla, V. C. (2008). *Un modèle stochastique pour la propagation du VIH/SIDAUne approche individuelle-centrée*. PhD thesis, Pau.

Kardas-sloma, L., Perozziello, A., Zahar, J.-R., Lescure, X., Yazdanpanah, Y., and Lucet, J. (2019). Transmission d'escherichia coli résistant aux β-lactamines (e. coli blse) dans la communauté : modélisation et évaluation de l'impact des interventions. *Médecine et Maladies Infectieuses*, 49(4) :S30–S31.

Kedowide Mevo Guezo, C. G. (2011). *SIG et analyse multicritère pour l'aide à la décision en agriculture urbaine dans les pays en développement, cas de Ouagadougou au Burkina Faso*. PhD thesis, Paris 8.

Kermack, W. O. and McKendrick, A. G. (1927). A contribution to the mathematical theory of epidemics. *Proceedings of the royal society of london. Series A, Containing papers of a mathematical and physical character*, 115(772) :700–721.

Kherdekar, V. A. and Metkewar, P. S. (2016). A technical comprehensive survey of etl tools. *International Journal of Applied Engineering Research*, 11(4) :2557–2559.

Khuu, M.-T. (2004). Génération d'un simulateur stochastique guidée par la description du système.

Kleinau, A., Stupak, E., Mörth, E., Garrison, L. A., Mittenentzwei, S., Smit, N. N., Lawonn, K., Bruckner, S., Gutberlet, M., Preim, B., and Meuschke, M. (2022). Is there a Tornado in Alex's Blood Flow ? A Case Study for Narrative Medical Visualization. In Raidou, R. G., Sommer, B., Kuhlen, T. W., Krone, M., Schultz, T., and Wu, H.-Y.,

editors, *Eurographics Workshop on Visual Computing for Biology and Medicine*. The Eurographics Association.

Kombila, U., Nzengue, E. E., Kinga, A., Mackanga, J., Mounguengui, D., Mbaye, F., Ba, J. I., and Boguikouma, J. (2017). Place de la tuberculose dans les péricardites aiguës en milieu hospitalier gabonais [tuberculosis and effusive pericarditidisin african hospital, libreville]. *Titre Page*, page 3.

Komorowski, M., Marshall, D. C., Salciccioli, J. D., and Crutain, Y. (2016). Exploratory data analysis. *Secondary analysis of electronic health records*, pages 185–203.

Koumamba, A. P., Lipenguet, G. M., Mbenga, R. O., Bisvigou, U. J., Assoum-Mve, F. U. A., Effame, Y. P., Tsokati, J. D., Nka, E. A., Djali, O. L., Ngoungou, B. E., et al. (2020). État des lieux du système d'information sanitaire du gabon. *Sante Publique*, 32(4) :407–417.

Kumar, V. A., Anandhi, M. D., Gopika, T., Devi, V. S., and Thenmozhi, S. (2019). Reliable data integration using talend.

Lacambre, A. (2001). *Aléas et risques naturels en montagne : apports et limites d'un Système d'Information Géographique (SIG) : application au haut bassin versant du Drac (Hautes-Alpes, France)*. PhD thesis, Paris 4.

Lacoste, O. (2006). L'usage des sig à l'observatoire régional de la santé du nord-pas-de-calais. *Géographes associés*, 30(1) :69–74.

Lambert, N. (2021). Cartographier la covid-19 : quelles narrations ? *Revue francophone sur la santé et les territoires*.

Lang, L. (2000). Gis for health organizations.

Larbani, B., Terniche, M., Taright, S., and Makhloufi, M. (2017). La prise en charge de la tuberculose pulmonaire dans une unité de contrôle de la tuberculose d'alger. *Revue des Maladies Respiratoires*, 34 :A230.

Larkey, L. K. and Gonzalez, J. (2007). Storytelling for promoting colorectal cancer prevention and early detection among latinos. *Patient education and counseling*, 67(3) :272–278.

Lee, B., Riche, N. H., Isenberg, P., and Carpendale, S. (2015). More than telling a story : Transforming data into visually shared stories. *IEEE Computer Graphics and Applications*, 35(5).

Leroi, E., Favre, J.-L., and Rezig, S. (2001). Cartographie de l'aléa mouvements de terrain par analyse statistique sous sig. *Revue française de géotechnique*, (95-96) :155–163.

Liu, C., Talaei-Khoei, A., Zowghi, D., and Daniel, J. (2017). Data completeness in healthcare : a literature survey. *Pacific Asia Journal of the Association for Information Systems*, 9(2) :5.

Lourenço, C., Tatem, A. J., Atkinson, P. M., Cohen, J. M., Pindolia, D., Bhavnani, D., and Le Menach, A. (2019). Strengthening surveillance systems for malaria elimination : a global landscaping of system performance, 2015–2017. *Malaria journal*, 18(1) :1–11.

Lousa, A., Pedrosa, I., and Bernardino, J. (2019). Evaluation and analysis of business intelligence data visualization tools. In *2019 14th Iberian Conference on Information Systems and Technologies (CISTI)*, pages 1–6. IEEE.

Majerovich, J., Fernandes, L., and Varia, M. (2017). Évaluation de la surveillance de l'infection tuberculeuse latente dans la région de peel, en ontario, 2010 à 2014. *RMTC*, 43 :5.

Mancini, F., Ceppi, C., and Ritrovato, G. (2010). Gis and statistical analysis for landslide susceptibility mapping in the daunia area, italy. *Natural Hazards and Earth System Sciences*, 10(9) :1851–1864.

Mate, K. S., Bennett, B., Mphatswe, W., Barker, P., and Rollins, N. (2009). Challenges for routine health system data management in a large public programme to prevent mother-to-child hiv transmission in south africa. *PloS one*, 4(5) :e5483.

Mboumba Sambo, Y. (2018). Lutte contre la consommation de la drogue en milieu scolaire au gabon : Cas du lycée paul idjendje gondjout de la commune de libreville.

McCall, B., Shallcross, L., Wilson, M., Fuller, C., and Hayward, A. (2019). Storytelling as a research tool and intervention around public health perceptions and behaviour : a protocol for a systematic narrative review. *BMJ open*, 9(12) :e030597.

McHugh, R., Roche, S., and Bédard, Y. (2007). Vers une solution solap comme outil participatif. In *Comptesrendus de la Conférence québéco-française pour le développement de la géomatique-CQFD*, volume 20.

McLafferty, S. L. (2003). Gis and health care. *Annual review of public health*, 24(1) :25–42.

Melki, B., Saad, S. B., Daghfous, H., Khelifa, M. B., and Tritar, F. (2015). Forme grave de la tuberculose : le pyopneumothorax tuberculeux. *Revue des Maladies Respiratoires*, 32 :A233.

Michel, F. (2006). Le modele irm4s : le principe influence/réaction pour la simulation de systemes multi-agents. In *Journées Francophones sur les Systèmes Multi-Agents*.

MinSanté (2020). Annuaire stastitiques saniatires 2020.

MPEAT and DGSEE (1992). Enquête budget consommation.

Nankep, N. et al. (2018). *Modélisation stochastique de systemes biologiques multi-échelles et inhomogenes en espace*. PhD thesis, Rennes, École normale supérieure.

Naoum, L. (2006). *Un modèle multidimensionnel pour un processus d'analyse en ligne de résumés flous*. Theses, Université de Nantes.

Natrella, M. (2010). Nist/sematech e-handbook of statistical methods, 2010. *URL : http ://goo. gl/8ElJs*.

Niang, S., ABDALLAHI, E., THIAM, K., MBAYE, F. B. R., CISSE, M., DIENG, A., BADIANE, N. T., et al. (2018). Aspects épidémiologiques, diagnostiques et évolutifs de la tuberculose pulmonaire à microscopie positive au district sanitaire de saint-louis. *Revue Africaine de Médecine Interne*, 5(2) :65–69.

Nikiforova, A. (2020). Definition and evaluation of data quality : User-oriented data object-driven approach to data quality assessment. *Baltic Journal of Modern Computing*, 8(3) :391–432.

Njeru, J. W., Patten, C. A., Hanza, M. M., Brockman, T. A., Ridgeway, J. L., Weis, J. A., Clark, M. M., Goodson, M., Osman, A., Porraz-Capetillo, G., et al. (2015). Stories for change : development of a diabetes digital storytelling intervention for refugees and immigrants to minnesota using qualitative methods. *BMC public health*, 15(1) :1–11.

Noah, N. (2006). *Controlling communicable disease*. McGraw-Hill Education (UK).

OMS (1966). Rapport sur les discussions techniques tenues à la dix-neuvième assemblée de la santé" l'établissement et l'utilisation des statistiques sanitaires dans les services nationaux et locaux de la santé". Technical report, Organisation mondiale de la Santé.

OMS (2020). Rapport sur la tuberculose dans le monde 2020 : résumé d'orientation.

Ondzigue Mbenga, R., Peralta, V., Devogele, T., Maghendji, S., and Ngoungou, E. B. (2021). Processus de narration de données en intelligence épidémique avec application à la pandémie de tuberculose au gabon. In *8e Journées Camerounaises d'Informatique Médicale*.

Ondzigue Mbenga, R., Peralta, V., Devogele, T., Outa, F. E., Nzondo, S. M., and Ngoungou, E. B. (2022). A data narrative about tuberculosis pandemic in gabon. In Ramanath, M. and Palpanas, T., editors, *Proceedings of the Workshops of the EDBT/ICDT 2022 Joint Conference, Edinburgh, UK, March 29, 2022*, volume 3135 of *CEUR Workshop Proceedings*. CEUR-WS.org.

Othman, A. (2016). *Simulation multi-agent de l'information des voyageurs dans les transports en commun*. PhD thesis, Université Paris-Est.

Palussière, M., Calavas, D., and Bronner, A. (2013). Evaluation de la qualité des données collectées dans le cadre du dispositif de déclaration obligatoire des avortements chez les bovins en france. *Bull. Epid. Santé Anim. Alim*, 58 :17–20.

Pareschi, M., Cavarra, L., Favalli, M., Giannini, F., and Meriggi, A. (2000). Gis and volcanic risk management. In *Natural hazards*, pages 361–379. Springer.

Passouant, M., Bélières, J.-F., Samaké, O., et al. (2000). Sig et suivi-évaluation de l'agriculture irriguée dans le delta du sénégal. *Science et changements planétaires/Sécheresse*, 11(2) :101–8.

Peddireddy, A. S., Xie, D., Patil, P., Wilson, M. L., Machi, D., Venkatramanan, S., Klahn, B., Porebski, P., Bhattacharya, P., Dumbre, S., et al. (2020). From 5vs to 6cs : Operationalizing epidemic data management with covid-19 surveillance. In *2020 IEEE Int. Conf. on Big Data*, pages 1380–1387.

Pesty, S., Brassac, C., and Ferrent, P. (1997). Ancrer les agents cognitifs dans l'environnement. *Quinqueton, Thomas et Trousse, éds., Actes Des.*

PNLT (2014). *Plan stratégique de lutte contre la tuberculose au Gabon 2014-2018*.

Prats, C., Montañola-Sales, C., Gilabert-Navarro, J. F., Valls, J., Casanovas-Garcia, J., Vilaplana, C., Cardona, P.-J., and López, D. (2016). Individual-based modeling of tuberculosis in a user-friendly interface : understanding the epidemiological role of population heterogeneity in a city. *Frontiers in Microbiology*, page 1564.

Proulx, M.-J., Bédard, Y., Nadeau, M., Gosselin, P., and Lebel, G. (2002). Géomatique et santé environnementale : innovations résultant du projet icem/se. *Géomatique 2002*.

Pushkarev, V., Neumann, H., Varol, C., Talburt, J. R., et al. (2010). An overview of open source data quality tools. In *IKE*, pages 370–376.

Rakotomanana, F., Rabarijaona, L. P., Ratsitorahina, M., Cauchoix, B., Razafinimanana, J., Ratsirahonana, A., Boisier, P., and Aurégan, G. (2000). Profil des malades perdus de vue en cours de traitement dans le programme national de lutte contre la tuberculose à madagascar. *Cahiers d'études et de recherches francophones/Santé*, 9(4) :225–229.

Rebman, A. W., Aucott, J. N., Weinstein, E. R., Bechtold, K. T., Smith, K. C., and Leonard, L. (2017). Living in limbo : contested narratives of patients with chronic symptoms following lyme disease. *Qualitative Health Research*, 27(4) :534–546.

Rivest, S., Gignac, P., Charron, J., and Bédard, Y. (2004). Développement d'un système d'exploration spatio-temporelle interactive des données de la banque d'information corporative du ministère des transports du québec. In *Colloque Géomatique-Un choix stratégique*.

Roques, L., Klein, E., Papaix, J., and Soubeyrand, S. (2020). Modèle sir mécanistico-statistique pour l'estimation du nombre d'infectés et du taux de mortalité par covid-19. *Rapport de recherche, INRAE*.

Santos, R. S., Malheiros, S. M., Cavalheiro, S., and De Oliveira, J. P. (2013). A data mining system for providing analytical information on brain tumors to public health decision makers. *Computer methods and programs in biomedicine*, 109(3) :269–282.

Savel, T. G., Foldy, S., for Disease Control, C., Prevention, et al. (2012). The role of public health informatics in enhancing public health surveillance. *MMWR Surveill Summ*, 61(2) :20–24.

Scannapieco, M., Missier, P., and Batini, C. (2005). Data quality at a glance. *Datenbank-Spektrum*, 14(January) :6–14.

Scotch, M. and Parmanto, B. (2005). Sovat : Spatial olap visualization and analysis tool. In *Proceedings of the 38th Annual Hawaii International Conference on System Sciences*, pages 142b–142b. IEEE.

Segel, E. and Heer, J. (2010). Narrative visualization : Telling stories with data. *IEEE Trans. Vis. Comput. Graph.*, 16(6) :1139–1148.

Sharma, A., Bahl, S., Bagha, A. K., Javaid, M., Shukla, D. K., Haleem, A., et al. (2020). Multi-agent system applications to fight covid-19 pandemic. *Apollo Medicine*, 17(5) :41.

Siebert, J. (2011). *Approche multi-agent pour la multi-modélisation et le couplage de simulations : application à l'étude des influences entre le fonctionnement des réseaux ambiants et le comportement de leurs utilisateurs*. PhD thesis, Université Henri Poincaré-Nancy 1.

Sissoko, B. F., Pileta, K. G., Dembele, J., Toloba, Y., Morales, D. C., Soumaré, D., Ouattara, K., N'Diaye, A., Pascual, G., and Diallo, S. (2013). Etude des facteurs conduisant les tubercuileux bacilliferes a une consultation tardive. *Mali Santé Publique*, pages 104–107.

Souibgui, M., Atigui, F., Zammali, S., Cherfi, S., and Yahia, S. B. (2019). Data quality in etl process : A preliminary study. *Procedia Computer Science*, 159 :676–687.

Straetemans, M., Bierrenbach, A. L., Nagelkerke, N., Glaziou, P., and van der Werf, M. J. (2010). The effect of tuberculosis on mortality in hiv positive people : a meta-analysis. *PLoS One*, 5(12) :e15241.

Sylla, A., Marchou, B., Kassi, N., Ello, N., Aba, T., Kouakou, G., Mossou, C., Ehui, E., Eholié, S., and Biassagnéné, E. (2017). Co-infection tuberculose/vih : à propos de 717 cas suivis dans un service de maladies infectieuses en afrique subsaharienne. *Médecine et Maladies Infectieuses*, 47(4) :S137–S138.

Taillandier, P. (2019). Vers une meilleure intégration des dimensions spatiales, comportementales et participatives en simulation à base d'agents.

Taillandier, P., Grignard, A., Gaudou, B., and Drogoul, A. (2014). Des données géographiques à la simulation à base d'agents : application de la plate-forme gama. *Cybergeo : European Journal of Geography*.

Taravat, N. A. (2009). Applications of gis in health sciences.

Tékpa, G., Fikouma, V., Téngothi, R. M. M., de Dieu Longo, J., Woyengba, A. P. A., and Koffi, B. (2019). Aspects épidémiologiques et cliniques de la tuberculose en milieu hospitalier à bangui. *The Pan African Medical Journal*, 33.

Teste, O. (2000). *Modélisation et manipulation d'entrepôts de données complexes et historisées*. PhD thesis, Université Paul Sabatier-Toulouse III.

Thacker, S. B., Qualters, J. R., Lee, L. M., for Disease Control, C., Prevention, et al. (2012). Public health surveillance in the united states : evolution and challenges. *MMWR Surveill Summ*, 61(3) :3–9.

Tolou, H. (2014). L'efficacité du vaccin bcg plus large qu'on ne l'estimait auparavant.

Tran, A., Guis, H., Guernier, V., and Gerbier, G. (2009). Epidémiologie spatiale : les maladies vues du ciel.

Tsui, E. K. and Starecheski, A. (2018). Uses of oral history and digital storytelling in public health research and practice. *Public health*, 154 :24–30.

Varaine, F. and Rich, M. L. (2014). *Tuberculose : guide pratique à l'usage des médecins, infirmiers, techniciens de laboratoire et auxiliaires de santé*. Médecins sans frontières.

Vassiliadis, P. and Simitsis, A. (2009). Extraction, transformation, and loading. *Encyclopedia of Database Systems*, 10.

VIDAL (2020). Tuberculose pulmonaire : prise en charge. `https://www.vidal.fr/maladies/recommandations/tuberculose-pulmonaire-1694.html#prise-en-charge`.

Vila Guilera, J. (2017). Analysis and individual-based modelling of the tuberculosis epidemiology in barcelona. the role of age, gender and origin. B.S. thesis, Universitat Politècnica de Catalunya.

Villani, C. (2020). Épidémie de covid-19-point sur la modélisation épidémiologique pour estimer l'ampleur et le devenir de l'épidémie de covid-19. *office parlementaire d'évaluation des choix scientifiques et technologiques*, 30.

Vroh, J. B. B., Noufé, S., Tiembre, I., Bogui, T. Y., Lepri, N. A., Yohou, K. S., Walley-Goli, C., Tagliante-Saracino, J., et al. (2015). Qualité des données de vaccination chez les enfants de 0 à 11 mois en côte d'ivoire. *Santé publique*, 27(2) :257–264.

Waggoner, P. E., Horsfall, J. G., et al. (1969). Epidem : a simulator of plant disease written for a computer. *Bulletin. Connecticut Agricultural Experiment Station*, 698.

Wand, Y. and Wang, R. Y. (1996). Anchoring data quality dimensions in ontological foundations. *Communications of the ACM*, 39(11) :86–95.

Weiskopf, N. G. and Weng, C. (2013). Methods and dimensions of electronic health record data quality assessment : enabling reuse for clinical research. *Journal of the American Medical Informatics Association*, 20(1) :144–151.

WHO (2014). Définitions et cadre de notification pour la tuberculose-révision 2013.

WHO (2014). Early detection, assessment and response to acute public health events : implementation of early warning and response with a focus on event-based surveillance. Technical document.

WHO (2014). Early detection, assessment and response to acute public health events : implementation of early warning and response with a focus on event-based surveillance : interim version. Technical report, World Health Organization.

WHO (2020). Tuberculosis profile : Gabon.

Wilson, J. (1999). Local, national, and global applications of gis in agriculture. *Geographical information systems : Principles, techniques, management, and applications*, pages 981–998.

Younsi, F. Z. (2016). Mise en place d'un système d'information décisionnel pour le suivi et la prévention des epidémies. *PhD diss. University of Lyon.*

Younsi, F.-Z., Bounnekar, A., Hamdadou, D., and Boussaid, O. (2019). Integration of multiple regression model in an epidemiological decision support system. *International Journal of Information Technology & Decision Making*, 18(06) :1755–1783.

Zargayouna, M. (2021). *Modèles multi-agents de déplacement à base d'activités : Etat de l'art*. PhD thesis, Université gustave eiffel.

Zeigler, B. P., Kim, T. G., and Praehofer, H. (2000). *Theory of modeling and simulation*. Academic press.

Ziyadi, N., Touzeau, S., Bidot, C., Treuil, J.-P., and Hbid, M. L. (2008). Modèle individu-centré de transmission de la tremblante dans un troupeau ovin.

Annexe A

Publications

— R. Ondzigue Mbenga, A.P. Koumamba, G. Moukoumbi Lipenguet, C.O. Bagayoko 2019. Usage des Technologies de l'Information et de la Communication par les professionnels de la santé du Gabon : Étude de perception dans le cadre du projet e-Santé Gabon. Journal of Health Informatics in Africa (JHIA) ;

— Raymond Ondzigue Mbenga, Thomas Devogele, Sydney Maghendji, Veronika Peralta, Edgard Brice Ngoungou. 2019 Un système dinformation géographique décisionnel pour la surveillance épidémiologique de la tuberculose en Afrique subsaharienne : Cas du Gabon. Conférence internationale francophone sur l'analyse spatiale et la géomatique (SAGEO 2019)

— Raymond Ondzigue Mbenga, Veronika Peralta, Thomas Devogele, Faten EL Outa, Sydney Maghendji Nzondo, Edgard Brice Ngoungou 2021. Processus de narration de données en intelligence épidémique avec application à la pandémie de tuberculose au Gabon. Journées camerounaises d'informatique médicale 2021 (JCIM 2021) ;

— Raymond Ondzigue Mbenga, Veronika Peralta, Thomas Devogele, Faten El Outa, Sydney Maghendji Nzondo, and Edgard Brice Ngoungou. 2022. A data narrative about tuberculosis pandemic in Gabon. EDBT/ICDT 2022 Workshops : Data Analytics solutions for Real-LIfe APplications (DARLI-AP 2022) ;

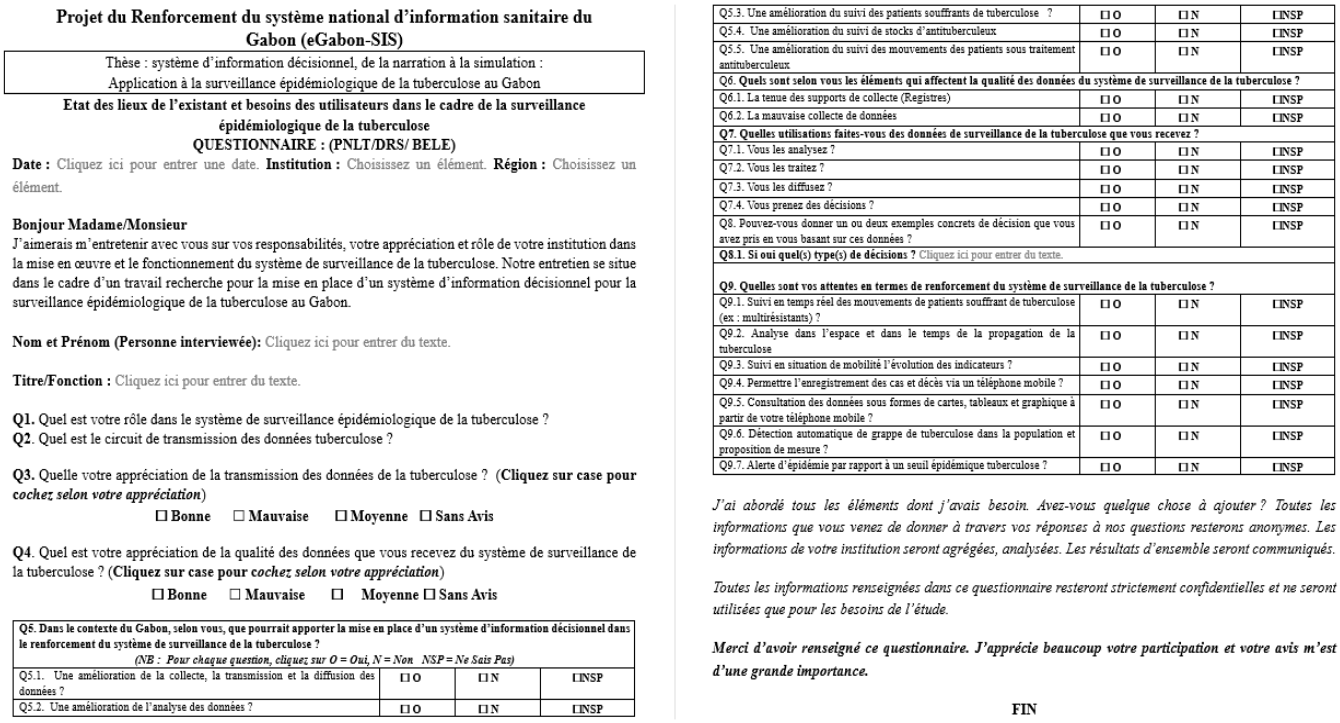
— Raymond Ondzigue Mbenga, Veronika Peralta, Edgard Brice Ngoungou, Sydney Maghendji Nzondo, Thomas Devogele, 2022. Narration de données en santé publique : cas de la tuberculose au Gabon. EDA 2022.

— Cheick Oumar Bagayoko, Jack Tchuente, Diakaridia Traoré, Gaetan Moukoumbi Lipenguet, Raymond Ondzigue Mbenga, Aimé Patrice Koumamba, Myriam Corille Ondjani, Olive Lea Ndjeli and Marie-Pierre Gagnon 2020. Implementation of a national electronic health information system in Gabon : a survey of healthcare providers perceptions. BMC Medical Informatics and Decision Making ;

— Aimé Patrice Koumamba, Gaetan Moukoumbi Lipenguet, Raymond Ondzigue Mbenga, Ulrich Jolhy Bisvigou, Fidéline Ursule Andeme Assoum-Mve, Yvon Patrice Effame, Jean Donatien Tsokati, Emmanuel Assoumou Nka, Olive Léa Djali, Brice Edgard Ngoungou, Gayo Diallo, Cheick Oumar Bagayoko 2020. Etat des lieux du système d'information sanitaire du Gabon. cairn.info ;

Annexe B

Questionnaire : État des lieux de l'existant et besoins des utilisateurs



Projet du Renforcement du système national d'information sanitaire du Gabon (eGabon-SIS)

Thèse : système d'information décisionnel, de la narration à la simulation :
Application à la surveillance épidémiologique de la tuberculose au Gabon

Etat des lieux de l'existant et besoins des utilisateurs dans le cadre de la surveillance épidémiologique de la tuberculose

QUESTIONNAIRE : (PNLT/DRS/ BELE)

Date : Cliquez ici pour entrer une date. **Institution :** Choisissez un élément. **Région :** Choisissez un élément.

Bonjour Madame/Monsieur

J'aimerais m'entretenir avec vous sur vos responsabilités, votre appréciation et rôle de votre institution dans la mise en œuvre et le fonctionnement du système de surveillance de la tuberculose. Notre entretien se situe dans le cadre d'un travail recherche pour la mise en place d'un système d'information décisionnel pour la surveillance épidémiologique de la tuberculose au Gabon.

Nom et Prénom (Personne interviewée): Cliquez ici pour entrer du texte.

Titre/Fonction : Cliquez ici pour entrer du texte.

Q1. Quel est votre rôle dans le système de surveillance épidémiologique de la tuberculose ?
Q2. Quel est le circuit de transmission des données tuberculose ?

Q3. Quelle votre appréciation de la transmission des données de la tuberculose ? (**Cliquez sur case pour *cochez selon votre appréciation***)

☐ **Bonne** ☐ **Mauvaise** ☐ **Moyenne** ☐ **Sans Avis**

Q4. Quel est votre appréciation de la qualité des données que vous recevez du système de surveillance de la tuberculose ? (**Cliquez sur case pour *cochez selon votre appréciation***)

☐ **Bonne** ☐ **Mauvaise** ☐ **Moyenne** ☐ **Sans Avis**

Q5. Dans le contexte du Gabon, selon vous, que pourrait apporter la mise en place d'un système d'information décisionnel dans le renforcement du système de surveillance de la tuberculose ? *(NB : Pour chaque question, cliquez sur O = Oui, N = Non NSP = Ne Sais Pas)*			
Q5.1. Une amélioration de la collecte, la transmission et la diffusion des données ?	☐ O	☐ N	☐NSP
Q5.2. Une amélioration de l'analyse des données ?	☐ O	☐ N	☐NSP
Q5.3. Une amélioration du suivi des patients souffrants de tuberculose ?	☐ O	☐ N	☐NSP
Q5.4. Une amélioration du suivi de stocks d'antituberculeux	☐ O	☐ N	☐NSP
Q5.5. Une amélioration du suivi des mouvements des patients sous traitement antituberculeux	☐ O	☐ N	☐NSP
Q6. Quels sont selon vous les éléments qui affectent la qualité des données du système de surveillance de la tuberculose ?			
Q6.1. La tenue des supports de collecte (Registres)	☐ O	☐ N	☐NSP
Q6.2. La mauvaise collecte de données	☐ O	☐ N	☐NSP
Q7. Quelles utilisations faites-vous des données de surveillance de la tuberculose que vous recevez ?			
Q7.1. Vous les analysez ?	☐ O	☐ N	☐NSP
Q7.2. Vous les traitez ?	☐ O	☐ N	☐NSP
Q7.3. Vous les diffusez ?	☐ O	☐ N	☐NSP
Q7.4. Vous prenez des décisions ?	☐ O	☐ N	☐NSP
Q8. Pouvez-vous donner un ou deux exemples concrets de décision que vous avez pris en vous basant sur ces données ?	☐ O	☐ N	☐NSP
Q8.1. Si oui quel(s) type(s) de décisions ? Cliquez ici pour entrer du texte.			
Q9. Quelles sont vos attentes en termes de renforcement du système de surveillance de la tuberculose ?			
Q9.1. Suivi en temps réel des mouvements de patients souffrant de tuberculose (ex : multirésistants) ?	☐ O	☐ N	☐NSP
Q9.2. Analyse dans l'espace et dans le temps de la propagation de la tuberculose	☐ O	☐ N	☐NSP
Q9.3. Suivi en situation de mobilité l'évolution des indicateurs ?	☐ O	☐ N	☐NSP
Q9.4. Permettre l'enregistrement des cas et décès via un téléphone mobile ?	☐ O	☐ N	☐NSP
Q9.5. Consultation des données sous formes de cartes, tableaux et graphique à partir de votre téléphone mobile ?	☐ O	☐ N	☐NSP
Q9.6. Détection automatique de grappe de tuberculose dans la population et proposition de mesure ?	☐ O	☐ N	☐NSP
Q9.7. Alerte d'épidémie par rapport à un seuil épidémique tuberculose ?	☐ O	☐ N	☐NSP

J'ai abordé tous les éléments dont j'avais besoin. Avez-vous quelque chose à ajouter ? Toutes les informations que vous venez de donner à travers vos réponses à nos questions resterons anonymes. Les informations de votre institution seront agrégées, analysées. Les résultats d'ensemble seront communiqués.

Toutes les informations renseignées dans ce questionnaire resteront strictement confidentielles et ne seront utilisées que pour les besoins de l'étude.

Merci d'avoir renseigné ce questionnaire. J'apprécie beaucoup votre participation et votre avis m'est d'une grande importance.

FIN

Figure 66 – Questionnaire (extrait) : états des lieux de l'existant et besoins des utilisateurs

Annexe C

Tableaux : résultats simulations

Table C.1 – Résultat simulation : scénario 1

Semestre	M	T	PDV	TA	G	ET	TI
1	3855	4013	1203	963	385	80	1557
2	3806	3962	1188	950	380	79	1537
3	3764	3918	1175	940	376	78	1520
4	3723	3875	1162	930	372	77	1503
5	3675	3826	1147	918	367	76	1484
6	3623	3771	1131	905	362	75	1463
7	3575	3721	1116	893	357	74	1443
8	3528	3672	1101	881	352	73	1424
9	3483	3626	1087	870	348	72	1406
10	3445	3586	1075	860	344	71	1391
11	3409	3549	1064	851	340	70	1377
12	3367	3505	1051	841	336	70	1359
13	3328	3464	1039	831	332	69	1344
14	3279	3413	1023	819	327	68	1324
15	3241	3374	1012	809	323	67	1309
16	3204	3335	1000	800	320	66	1293
17	3166	3296	988	791	316	65	1278
18	3130	3258	977	781	312	65	1264
19	3101	3228	968	774	309	64	1252
20	3068	3194	958	766	306	63	1239

TABLE C.2 – Résultat simulation : Nombre de malades, des PDV et des guéris sur 20 semestres après augmentation du taux d'individus mis en traitement de 51% à 70%

Semestre	M (T=51%)	M(T=70%)	PDV (T=51%)	PDV (T=70%)	G (T=51%)	G (T=70%)
1	3855	3855	1203	1203	385	385
2	3855	1956	1203	552	385	971
3	3821	1936	1193	546	381	962
4	3778	1913	1179	540	377	950
5	3728	1887	1164	532	372	937
6	3691	1863	1152	526	368	925
7	3650	1838	1139	519	364	913
8	3599	1812	1123	511	359	900
9	3558	1789	1111	505	355	888
10	3516	1770	1098	499	351	879
11	3477	1745	1085	492	347	867
12	3436	1721	1073	485	343	855
13	3394	1699	1059	479	339	844
14	3348	1678	1045	473	334	833
15	3305	1655	1032	467	330	822
16	3266	1635	1020	461	326	812
17	3220	1613	1005	455	321	801
18	3182	1594	993	450	317	792
19	3136	1574	979	444	313	782
20	3091	1557	965	439	308	773

TABLE C.3 – Résultat simulation : Nombre d'individus susceptibles et immunisés (augmentation du taux vaccinés immunisés de 70% à 80%)

Semestre	S (Vacc-Imm=70%)	S (Vacc-Imm=80%)	IMM (Vacc-Imm=70%)	(Vacc-Imm=80%)
1	559764	559764	79918	79918
2	554601	554182	85081	85500
3	549517	548761	90165	90921
4	544531	543371	95151	96311
5	539525	537962	100157	101720
6	534769	532734	104913	106948
7	529856	527518	109826	112164
8	525002	522591	114680	117091
9	520258	517593	119424	122089
10	515685	512743	123997	126939
11	511012	507978	128670	131704
12	506760	503302	132922	136380
13	502447	498646	137235	141036
14	498200	493990	141482	145692
15	494029	489414	145653	150268
16	489900	484938	149782	154744
17	485900	480731	153782	158951
18	481924	476468	157758	163214
19	478009	472333	161673	167349
20	474210	468217	165472	171465

TABLE C.4 – Résultat simulation : nombre d'individus malades (M), ayant achevé le traitement (TA) et guéris(G) sur 20 semestres après diminution du taux d'individus perdus de vue de 30% à 10%

Semestre	M (PDV=30%)	M (PDV=10%)	TA (T=30%)	TA (T=10%)	G (PDV=30%)	G (PDV=10%)
1	3855	3855	963	963	385	385
2	3821	3778	954	1730	381	692
3	3778	3731	943	1708	377	683
4	3728	3667	931	1679	372	671
5	3691	3613	922	1654	368	661
6	3650	3547	911	1624	364	649
7	3599	3494	899	1600	359	640
8	3558	3449	888	1579	355	631
9	3516	3401	878	1557	351	623
10	3477	3353	868	1535	347	614
11	3436	3315	858	1518	343	607
12	3394	3281	847	1502	339	601
13	3348	3233	836	1480	334	592
14	3305	3186	825	1459	330	583
15	3266	3149	816	1442	326	576
16	3220	3105	804	1422	321	568
17	3182	3062	794	1402	317	561
18	3136	3005	783	1376	313	550
19	3091	2964	772	1357	308	542
20	3069	2927	766	1340	306	536

Raymond ONDZIGUE MBENGA

Système d'information décisionnel, de la narration à la simulation : application à la surveillance épidémiologique de la tuberculose au Gabon

Résumé :

La tuberculose (TB) demeure un grave problème de santé publique au Gabon. En effet, selon l'Organisation Mondiale de la Santé (OMS), le pays est désormais compté parmi les 30 pays du monde à forte charge tuberculose. Malgré cette situation épidémiologique très inquiétante, les experts et autorités de la santé publique ne disposent pas aujourd'hui d'outils informatiques adaptés pour surveiller cette maladie. C'est dans ce contexte que nous nous intéressons dans cette thèse à l'élaboration d'un système d'aide à la décision pour le suivi, la surveillance, l'analyse et la simulation de politiques sanitaires de riposte pour aider les autorités sanitaires à lutter efficacement contre cette pandémie. La construction d'un tel système pose plusieurs problèmes de recherche, notamment (1) d'intégration des données hétérogènes (actuellement traitées manuellement), (2) la résolution de multiples problèmes de qualité, (3) l'analyse de données spatio-temporelles, (4) la restitution des indicateurs à un public d'experts en santé publique mais pas informatique et (5) la simulation de diverses politiques sanitaires. Pour répondre à ces enjeux, nous proposons un système décisionnel qui couple un système d'information géographique décisionnel de la tuberculose (SOLAP-TB) avec un système multi-agents de la tuberculose (SMA-TB). L'objectif est de donner aux experts et autorités de la santé publique la possibilité à travers un même outil, d'une part de surveiller l'évolution spatio-temporelle de cette pandémie via des tableaux de bord interactifs et dynamiques d'indicateurs puis d'autre part de simuler les politiques sanitaires de riposte via un modèle prédictif. En outre, afin de faciliter la compréhension de la situation épidémiologique de la TB aux experts de la santé publique qui sont des décideurs, nous proposons un processus de narration de données en intelligence épidémique (P-N-D-I-E) qui découle des bonnes pratiques en narration de données et en intelligence épidémique de l'OMS. Ensuite, en utilisant ce processus, nous avons produit une narration de données sur la pandémie de la tuberculose au Gabon qui présente les trouvailles extraites des analyses des données réalisées au niveau du SOLAP-TB. Finalement, le système décisionnel SOLAP-SMA de la TB que nous avons proposé est générique, c'est-à-dire qu'il peut être adapté à la surveillance de la TB dans d'autres pays. Par ailleurs, le processus de narration de données en intelligence épidémique peut être utilisé pour produire des narrations de données sur d'autres maladies.

Mots clés : SOLAP, SMA, Processus de narration de données en intelligence épidémique, tuberculose, Gabon

Summary :

Tuberculosis (TB) remains a serious public health problem in Gabon. According to the World Health Organization (WHO), the country is now counted among the 30 countries in the world with a high burden of TB. In spite of this very worrying epidemiological situation, experts and public health authorities do not currently have the appropriate computerized tools to monitor this disease. It is in this context that we are interested in the development of a decision support system for monitoring, surveillance, analysis and simulation of sanitary response policies to help health authorities to fight effectively against this pandemic. The construction of such a system poses several research challenges, including (1) the integration of heterogeneous data (currently processed manually), (2) the resolution of multiple quality issues, (3) the analysis of spatiotemporal data, (4) the restitution of indicators to an audience of public health but not computers experts and (5) the simulation of various health policies. To address these issues, we propose a decision-making system that couples a geographic information system for TB (SOLAP-TB) with a multi-agent TB system (SMA-TB). The objective is to give experts and public health authorities the possibility, through the same tool, to monitor the spatio-temporal evolution of this pandemic via interactive dashboards of indicators and then to simulate the sanitary policies of response via a predictive model. In addition, in order to facilitate the understanding of the TB epidemiological situation to public health experts who are decision makers, we propose a data storytelling process in epidemic intelligence that follows the WHO good practices in data storytelling and epidemic intelligence. Then, using this process, we produced a data narrative on the TB pandemic in Gabon that presents the findings extracted from the data analyses performed at the SOLAP-TB level. Finally, the SOLAP-SMA TB system we propose is generic, i.e., it can be adapted to TB surveillance in other countries. Furthermore, the epidemic intelligence data storytelling process can be used to produce data stories for other diseases.

Keywords : SOLAP, SMA, Data storytelling process in epidemic intelligence, tuberculosis, Gabon