2026

00

Bases statistiques et généralités

Neuchâtel 2026

Statistique du mouvement naturel de la population

Flash estimates

Schweizerische Eidgenossenschaft
Confédération suisse
Confederazione Svizzera
Confederaziun svizra

Département fédéral de l'intérieur DFI
Office fédéral de la statistique OFS

Rapports de méthodes de la section Méthodes statistiques

Les rapports de méthodes décrivent les méthodes mathématiques et statistiques à la base des résultats et des analyses de la statistique publique. Ils présentent également l'évaluation et le développement de nouvelles méthodes en vue d'une application future. Ces publications visent d'une part à documenter les méthodes utilisées ou envisagées dans un souci de transparence et de rigueur scientifique, et d'autre part à favoriser la collaboration avec le monde scientifique et universitaire.

Les résultats numériques présentés dans les rapports de méthodes illustrent les concepts mathématiques décrits, mais ne sont pas des résultats officiels des enquêtes concernées. De même, les méthodes réellement appliquées peuvent différer légèrement de celles décrites dans ces rapports.

Les rapports de méthodes sont disponibles sous forme électronique sur le site Internet de l'OFS.

Statistique du mouvement naturel de la population

Flash estimates

Rédaction Clément Chevalier, OFS
Djalel-Eddine Meskaldji, OFS
Éditeur Office fédéral de la statistique (OFS)

Neuchâtel 2026

Éditeur Office fédéral de la statistique (OFS)
www.bfs.admin.ch/fr

Renseignements Clément Chevalier, tél. +41 58 481 00 19
clement.chevalier@bfs.admin.ch

Rédaction Clément Chevalier, OFS; Djalel-Eddine Meskaldji, OFS

Contenu Section méthodes statistiques, OFS

Série Statistique de la Suisse

Domaine 00 Bases statistiques et généralités

Langue du texte original français

Mise en page Publishing et diffusion PUB, OFS

En ligne www.statistique.admin.ch/338-0096

Numéro OFS 338-0096

ISBN 978-3-303-00788-4

Distribution www.publicationsfederales.admin.ch
No d'art. 317.098.F

Copyright OFS, Neuchâtel 2026
La reproduction est autorisée si la source est mentionnée.

Table des matières

Préambule et Résumé **5**

1 Introduction **7**

2 Données, stratégie, préparation et intégrité **8**
2.1 Données disponibles . . . 8
2.2 Stratégie pour produire les flash estimates . . . 8
2.3 Outliers et imputations . . . 9

3 Modélisation **12**
3.1 Présentation générale des modèles REG-ARIMA . . . 12
3.2 Choisir un modèle REG-ARIMA . . . 13
3.2.1 Transformation préliminaire et outliers . . . 13
3.2.2 Effets calendrier . . . 14
3.2.3 Choix du modèle ARIMA saisonnier . . . 15
3.2.4 Vue synthétique de la procédure du choix de modèle . . . 16
3.3 Modélisation des mariages . . . 17
3.3.1 Imputations . . . 17
3.3.2 Outliers . . . 19
3.3.3 Effets calendrier . . . 22
3.3.4 Modèle ARIMA et conclusions . . . 23
3.4 Modélisation des divorces . . . 24
3.4.1 Imputations . . . 24
3.4.2 Outliers . . . 25
3.4.3 Effets calendrier . . . 26
3.4.4 Modèle ARIMA et conclusions . . . 26
3.5 Modélisation des naissances . . . 27
3.5.1 Imputations . . . 27
3.5.2 Outliers . . . 27
3.5.3 Effets calendrier . . . 29
3.5.4 Modèle ARIMA et conclusions . . . 29
3.6 Modélisation des décès . . . 30
3.6.1 Imputations . . . 30
3.6.2 Outliers . . . 31
3.6.3 Effets calendrier . . . 32
3.6.4 Modèle ARIMA et conclusions . . . 33
3.7 Synthèse sur les modèles et améliorations possibles . . . 33

4 Performance des modèles en simulation **35**
4.1 Description des simulations effectuées . . . 35
4.2 Performances pour les mariages . . . 36
4.3 Performances pour les divorces . . . 37
4.4 Performances pour les naissances . . . 38
4.5 Performances pour les décès . . . 38

5 Travail de production des flash estimates **39**

6 Conclusions et perspectives **42**

Bibliographie **43**

Annexe **44**

A Exemple d'analyse de révision pour le modèle des données provisoires des mariages **44**

Rapports de méthodes de la section méthodes statistiques de l'OFS **45**

Préambule

Dans le cadre de la statistique du mouvement naturel de la population, l'Office Fédéral de la Statistique (OFS) publie chaque année, vers le milieu du mois de juin, des données définitives sur le nombre de mariages, divorces, naissances et décès (MDND) survenus au cours de l'année révolue. Ainsi, le 19 juin 2025, l'OFS a publié un communiqué de presse et des résultats relatifs à l'ensemble de l'année 2024. Le délai pour publier des données relatives à l'année révolue s'explique par la disponibilité tardive des données définitives. Ces données n'étant pas prêtes au mois de janvier ou février, il faut ainsi attendre le mois de juin pour publier.

Si les données définitives des MDND ne sont pas disponibles avant juin, l'OFS dispose en revanche, dès le début d'année, de données provisoires. Ces données provisoires des MDND sont influencées par des décalages temporels dans les déclarations et ne constituent pas, en l'état, une prévision satisfaisante des valeurs définitives. Dans ce contexte, la section des méthodes statistiques (METH) de l'OFS a été sollicitée pour travailler sur des "flash estimates", c'est-à-dire des estimations des données définitives des MDND qui seraient publiées plus tôt et qui se baseraient sur l'information disponible lors de cette publication, notamment les données provisoires.

Les estimations produites dans ce rapport se basent sur une modélisation REG-ARIMA de données mensuelles, c'est-à-dire sur une régression linéaire qui prend en compte d'éventuels outliers et effets calendrier ainsi qu'un modèle ARIMA saisonnier pour traiter les corrélations temporelles. Des prédictions sont effectuées pour chaque mois de l'année révolue, et l'agrégation de ces prédictions permet d'obtenir des flash estimates pour l'année considérée. Les simulations montrent des résultats encourageants avec des erreurs relatives de l'ordre de 2% pour les naissances, décès ou divorces, et légèrement supérieures à 2% pour les mariages. Les auteurs remercient les membres de la section METH pour la relecture attentive de ce rapport, Daniel Assoulin, Daniel Kilchmann et Jean-Pierre Renfer. Les auteurs remercient également les collaborateurs de la section Démographie & Migration (DEM) qui ont relu le rapport et travaillé avec METH à l'élaboration des méthodes, Stephan Gysi et Céline Schmid.

Résumé

La statistique du mouvement naturel de la population est publiée chaque année par l'Office Fédéral de la Statistique. Le nombre total de mariages, divorces, naissances et décès au cours de l'année écoulée est publié au mois de juin de l'année suivante. L'objectif de ce travail est d'étudier la possibilité et la pertinence de publier des "flash estimates" plusieurs mois avant la publication finale. Les flash estimates sont effectuées à partir des données provisoires et se basent sur une modélisation REG-ARIMA de séries temporelles. Celle-ci n'offre cependant pas de garantie théorique sur la précision des estimations. En particulier des évènements importants peuvent causer des changements dans le comportement des séries et ainsi affecter la précision des résultats, du fait d'une violation de l'hypothèse de stationnarité sous-jacente aux modèles. Ceci est souligné dans le rapport à travers des exemples. Le choix de paramètres des modèles est décrit en détail et une analyse des performances, sous la forme d'un backtest, est fournie. Les backtests montrent que l'utilisation de données provisoires, disponibles avant les données définitives, et d'une modélisation adaptée permet d'obtenir des écarts de l'ordre de 2% entre les flash estimates et les chiffres définitifs.

1 Introduction

La statistique du mouvement naturel de la population est publiée chaque année par l'Office Fédéral de la Statistique (OFS) et fournit des informations sur le nombre de mariages, divorces, naissances et décès (MDND) au cours de l'année écoulée. Les publications sont disponibles généralement vers le mois de juin sur le site internet de l'OFS [1]. Ainsi, en juin 2026, l'OFS publie les nombres de MDND de l'année 2025.

Les raisons principales du délai relativement important pour disposer de données définitives de l'année révolue sont l'enregistrement parfois tardif des MDND aux offices de l'état civil des cantons ainsi que le traitement de ces données. L'enregistrement, la transmission et le traitement des données à l'OFS peuvent ainsi prendre plusieurs mois.

Si certains enregistrements de MDND sont tardifs, la majorité de ces enregistrements se font sur un délai relativement court. Il est ainsi possible de disposer de données mensuelles dites "provisoires" de MDND, qui peuvent être vues comme les données définitives sans les enregistrements tardifs. Des séries temporelles mensuelles peuvent être construites sur la base des données provisoires et définitives. Il s'agit de séries de nombres entiers qui partagent entre elles des liens évidents. Chaque valeur provisoire, pour un mois et une année donnée, est strictement plus petite que la valeur définitive correspondante. Au début d'une année donnée, les données provisoires pour l'année révolue sont généralement disponibles pour les mois de janvier à novembre. Les données provisoires de décembre sont livrées un mois plus tard.

Les séries temporelles basées sur les données définitives ou provisoires présentent des caractéristiques typiques, notamment une saisonnalité, des effets calendrier et des corrélations temporelles. De plus, compte tenu des explications précédentes, les données définitives à prédire peuvent être vues comme la somme entre les données provisoires et le nombre d'enregistrements tardifs. Dans cette logique, chercher à prédire les données définitives revient à prédire une somme entre des données provisoires, qui sont déjà disponibles en début d'année jusqu'au mois de novembre, et le nombre d'enregistrements tardifs. Ce dernier est égal à la différence entre les données mensuelles définitives et provisoires et les analyses montrent que cette différence est une série temporelle avec les caractéristiques décrites précédemment. Ce rapport décrit ainsi des choix pour deux modèles différents :

- Un modèle qui vise à prédire le résultat provisoire manquant, celui du mois de décembre de l'année révolue, au début d'une année donnée. Le modèle se base sur les données provisoires disponibles jusqu'au mois de novembre de l'année révolue.
- Un modèle qui vise à prédire le nombre d'enregistrements tardifs, c'est-à-dire la différence entre les données définitives et provisoires.

Cette approche est effectuée indépendamment pour les mariages, divorces, naissances et décès ce qui revient à construire 8 modèles au total. Tous les modèles utilisés ici sont des modèles de régression avec erreurs ARIMA, REG-ARIMA, comme précisé dans ce rapport. L'objectif n'est pas de remplacer les statistiques définitives publiées par l'OFS, mais de fournir une estimation plusieurs mois plus tôt. Les modèles ARIMA, "autoregressive integrated moving average", sont courants pour traiter les séries temporelles et sont décrits par exemple dans Brockwell & Davis [2009].

Ce rapport est organisé comme suit. La section 2 présente les données disponibles et les traitements liés à leur intégrité. La stratégie générale adoptée pour produire les flash estimates est aussi décrite. La section 3 présente les modèles REG-ARIMA utilisés et l'ensemble des aspects liés aux choix des différents paramètres. Chacun des 8 modèles est décrit dans cette

1. https://www.bfs.admin.ch/bfs/fr/home/statistiques/population.html

section, avec des justifications et des perspectives. La section 4 évalue la précision des flash estimates calculées à l'aide de ces modèles via un backtest [2] . Finalement, la section 5 présente le processus de production des flash estimates et les questions liées à l'arrivée de nouvelles données. Les perspectives d'amélioration des modèles et des performances sont évoquées en conclusion.

2 Données, stratégie, préparation et intégrité

2.1 Données disponibles

La section Démographie & Migration (DEM) de l'OFS a livré à la section des méthodes statistiques (METH) des données mensuelles de la statistique du mouvement naturel de la population en Suisse pour les quatre variables d'intérêt suivantes : mariages, divorces, naissances et décès (MDND). Pour chaque variable, deux séries temporelles sont disponibles : les données provisoires et les données définitives. Les valeurs des séries correspondent au nombre d'évènements (mariage, décès, naissance ou divorce) enregistrés en Suisse au cours du mois considéré et sont donc des nombres entiers. Comme expliqué précédemment, les données définitives correspondent aux valeurs consolidées après validation complète des déclarations administratives. Elles sont considérées comme les valeurs de référence et sont publiées chaque année par l'OFS. Les données provisoires ne sont pas publiées et correspondent aux déclarations enregistrées à un moment donné. En raison des délais dans les processus administratifs et de transmission des informations, les données provisoires sous-estiment toujours les valeurs définitives. Le domaine d'étude des évènements MDND est toujours la Suisse entière et le détail par canton ou autre subdivision n'est pas étudié dans ce rapport. Pour les mariages, naissances et décès, des données provisoires et définitives mensuelles sont disponibles sur une période allant de janvier 2011 à décembre 2024, soit 14 années de données mensuelles ou 168 mois. Pour les divorces, les données vont de janvier 2013 à décembre 2024, soit 12 années de données mensuelles ou 144 mois.

Certaines données provisoires livrées sont manquantes ou incomplètes et un contrôle de l'intégrité des données est nécessaire. Celui-ci est décrit en section 2.3. Les contrôles peuvent mener à un traitement des données brutes par imputation. C'est avec ces données corrigées que les estimations de modèles décrites en section 3 sont effectuées.

2.2 Stratégie pour produire les flash estimates

Soit a une année pour laquelle le but est d'estimer sur toute la Suisse le nombre total d'un évènement E donné : mariage, divorce, naissance ou décès. À la fin du mois de janvier de l'année $a+1$, l'OFS dispose des données suivantes :

- La série temporelle mensuelle $\text{prov}_{E,t}$ des données provisoires pour l'évènement E considéré, jusqu'au mois de novembre de l'année a,
- la série temporelle mensuelle $\text{def}_{E,t}$ des données définitives pour l'évènement E considéré, jusqu'au mois de décembre de l'année $a-1$.

Les séries temporelles considérées sont supposées être déjà traitées, via une imputation des données manquantes ou incomplètes. Ce processus de préparation de ces données est décrit en section 2.3. L'objectif est de prédire les valeurs de la série des données définitives sur les 12

2. Le backtest est une procédure de validation qui consiste à simuler l'utilisation des modèles en conditions réelles en entraînant ces derniers uniquement sur les données historiques jusqu'à une date donnée, par exemple décembre 2019, pour prédire les observations ultérieures déjà disponibles, par exemple l'année 2020.

mois de l'année a. L'agrégation de ces 12 prédictions permet d'obtenir une flash estimate sur le nombre total de E en Suisse au cours de l'année a.

Comme évoqué en section 1, la stratégie pour produire les flash estimates repose sur le constat que les données définitives sont in fine la somme de deux séries temporelles distinctes : les données provisoires et le nombre d'enregistrements tardifs. Prédire les données définitives revient ainsi à prédire au mieux chacune de ces deux composantes sur toute l'année a considérée.

1. Pour les données provisoires, $\text{prov}_{E,t}$, seule la valeur de décembre de l'année a est manquante au moment où les estimations sont produites. Il s'agit donc de prédire au mieux cette valeur, ce qui revient finalement à prédire la prochaine valeur d'une série temporelle mensuelle.
2. Pour les enregistrements tardifs, il est possible de noter que ceux-ci sont égaux à la différence $\text{diff}_{E,t} := \text{def}_{E,t} - \text{prov}_{E,t}$. Cette différence est une série temporelle qui est connue jusqu'au mois de décembre de l'année $a-1$ au moment où les estimations sont produites. Il s'agit donc ici de prédire cette série sur toute l'année a à partir de données disponibles jusqu'en fin d'année $a-1$, c'est-à-dire d'estimer les 12 prochaines valeurs de la série.

Les modèles ARIMA sont des candidats naturels pour traiter ce type de questions [Brockwell & Davis, 2009]. Comme les séries présentent parfois des effets calendrier et des outliers, voir section 3, les modèles choisis sont des modèles REG-ARIMA qui représentent la série étudiée comme la somme ou le produit d'une série prenant en compte les effets calendrier et outliers et d'une série temporelle suivant un modèle ARIMA. Une présentation générale des modèles REG-ARIMA et de l'estimation des paramètres est donnée dans le manuel de X13-ARIMA-SEATS, U.S. Census Bureau [2025], section 4. L'outil X13-ARIMA-SEATS est développé par le bureau du recensement des États-Unis et permet d'estimer des modèles REG-ARIMA. Dans ce rapport, cet outil est utilisé directement depuis une interface R [R Core Team, 2021] grâce au package "seasonal" [Sax & Eddelbuettel, 2018].

La section 3 décrit ainsi les choix de modèles REG-ARIMA pour les séries $\text{prov}_{E,t}$ et $\text{diff}_{E,t}$ pour les 4 types d'évènements E : mariages, divorces, naissances et décès. Un total de 8 modèles sont ainsi décrits. Avant cela, les données utilisées dans le cadre de ce travail sont contrôlées et éventuellement modifiées. Cette étape de préparation des données est décrite dans la section suivante.

2.3 Outliers et imputations

Les données livrées par la section DEM à la section METH peuvent présenter des valeurs inattendues, classiquement appelées "outliers". Des outliers sont des valeurs inattendues dans le sens où elles tombent loin des prédictions qui auraient été effectuées avec les modèles développés ici. Par exemple, le nombre de décès en novembre 2020, très affecté par la seconde vague de la pandémie de COVID-19, peut être considéré comme un outlier puisque le modèle seul ne serait pas capable de prédire ce nombre élevé de décès sur la base des données du passé.

Il est crucial dans ce travail de faire la distinction entre les outliers et les données manquantes ou incomplètes. Un outlier est une valeur inattendue sur la base du seul modèle à disposition, mais ce n'est pas pour autant une valeur fausse ou incomplète. Les modèles ARIMA sont connus pour mal gérer les outliers, ceux-ci perturbant fortement l'estimation des paramètres et donc l'ensemble du modèle. L'outil X-13ARIMA-SEATS est capable de gérer les outliers dans les séries temporelles grâce à la partie REG des modèles REG-ARIMA. Les solutions proposées par

X13-ARIMA-SEATS sont aujourd'hui largement répandues et sont utilisées dans ce travail. Le traitement des outliers est détaillé pour les mariages, divorces, naissances et décès en sections 3.3.2, 3.4.2, 3.5.2, 3.6.2.

À la différence des outliers, les séries temporelles peuvent avoir des données manquantes, voire même des données incomplètes. Ceci peut se produire en cas de problèmes dans les enregistrements provisoires des MDND. Dans ce cas-là, contrairement aux outliers, c'est l'intégrité des données elles-mêmes qui est en jeu et une imputation de la valeur concernée est nécessaire. Dans le jeu de données livrées par la section DEM, les données définitives, c'est-à-dire les données officiellement publiées par l'OFS, contiennent des outliers mais pas de données manquantes ou incomplètes. En revanche, pour les données provisoires, les constats suivants sont effectués :

- pour les mariages, naissances et décès, les données de décembre 2011, décembre 2012 et décembre 2013 sont manquantes. Par ailleurs, les données provisoires de novembre et décembre 2024 pour les décès semblent poser problème. Ceci est constaté en notant que les données définitives sont conformes à ce qui serait attendu pour un mois de novembre ou décembre alors que les données provisoires donnent des valeurs sensiblement inférieures à l'attendu, si bien que X13-ARIMA-SEATS détecte automatiquement des outliers sur la série $\text{diff}_{E,t}$ des décès en novembre et décembre 2024, sans qu'une interprétation crédible puisse être obtenue.
- Pour les divorces, la valeur de décembre 2013 est manquante. De plus, une analyse approfondie, qui fait également appel à une détection automatique par X13-ARIMA-SEATS, montre que les valeurs d'octobre et novembre 2017 sont incomplètes, en raison d'un problème de transmission des données.

Comme les 4 séries MDND n'ont aucune valeur manquante à partir de janvier 2014, une possibilité pour traiter le problème des données manquantes de décembre 2011, 2012 et 2013 serait de baser les analyses sur une série de données plus courte, commençant en janvier 2014. Le principal inconvénient de cette approche est de se priver de 3 années de données "presque complètes" pour les mariages, naissances et décès et d'une année pour les divorces. Les données sont effectivement presque complètes car il ne manque aucune valeur définitive et il manque simplement les données provisoires de décembre pour les années 2011, 2012 et 2013. Travailler sur des données allant de janvier 2014 à décembre 2024 revient à travailler sur 11 années de données mensuelles, soit 132 mois. Cette approche n'a pas été retenue pour les deux raisons suivantes :

1. Nombre de paramètres des modèles : les modèles REG-ARIMA obtenus ont un nombre de paramètres relativement important, voir section 3.7. Par exemple, pour les mariages, les modèles peuvent atteindre 20 paramètres environ. Pour obtenir des modèles stables dans le temps et éviter l'overfitting, il est nécessaire de baser l'estimation de ces paramètres sur des jeux de données les plus grands possibles. En particulier, le nombre d'observations doit rester grand face au nombre de paramètres.
2. Backtests : les performances des modèles obtenus sont testées en section 4. L'objectif est d'estimer la précision des flash estimates, c'est-à-dire l'ordre de grandeur de la différence entre les flash estimates effectuées fin janvier/début février et les données définitives disponibles en juin. Ces tests consistent en des prédictions des données définitives pour les années 2019 à 2024 à partir des données du passé. En particulier, les prédictions des données définitives de l'année 2019 se basent sur des données provisoires qui se terminent en novembre 2019 et des données définitives qui vont jusqu'à décembre 2018, voir section 2.2 pour un rappel sur la stratégie de prédiction. La série des données définitives se terminant en décembre 2018 doit comporter un nombre d'observations suffisamment

important pour estimer un modèle de manière stable. Une suppression des données pour les années 2011 à 2013 réduirait cette série à seulement 5 années de données mensuelles, ou 60 mois, ce qui est insuffisant compte tenu du nombre de paramètres des modèles mentionné précédemment.

Pour ces raisons, il est décidé de procéder à des imputations sur les données provisoires manquantes de sorte à avoir des données complètes pour les années 2011, 2012 et 2013. Les valeurs provisoires d'octobre et novembre 2017 pour les divorces sont aussi remplacées par des valeurs imputées, tout comme les valeurs provisoires de novembre et décembre 2024 pour les décès. L'outil X13-ARIMA-SEATS peut être utilisé pour effectuer cette tâche. La méthode consiste à remplacer les valeurs manquantes ou incomplètes par des valeurs très hautes, par exemple 99'999 et de déclarer ces valeurs comme des outliers dans le modèle REG-ARIMA. X13-ARIMA-SEATS est alors capable de produire une série ajustée des outliers qui conserve les caractéristiques de la série initiale (tendance, saisonnalité, effets calendrier). Plus spécifiquement, la série X_t de départ peut être décomposée en un produit

$$X_t := X'_t \, O_t \tag{1}$$

où O_t est une série mensuelle qui prend en compte les corrections d'outliers. Si aucune correction n'est effectuée à un instant t donné, O_t prend la valeur 1. La série ajustée de ses outliers X'_t est récupérée en calculant simplement le ratio X_t/O_t. Pour les mariages et naissances, la valeur de X'_t en décembre 2011, 2012 et 2013 sert à imputer les valeurs manquantes. Pour les divorces, ce sont les valeurs de décembre 2013, octobre 2017 et novembre 2017 qui sont prises en compte pour gérer la valeur manquante et les deux données incomplètes. Pour les décès, ce sont les valeurs de décembre 2011, 2012 et 2013 ainsi que novembre et décembre 2024 qui sont prises en compte.

L'imputation des données manquantes ou incomplètes via X13-ARIMA-SEATS nécessite un choix de modèle, ce qui est précisément l'objet de ce travail. La procédure se fait en deux étapes. Premièrement, un premier modèle est choisi sur la base des données disponibles. Ce modèle produit une première imputation grâce à l'équation (1). Deuxièmement, un choix final de modèle est fait sur les données complétées. Dans le cas où le modèle final est différent du modèle d'imputation les imputations sont corrigées en utilisant le nouveau modèle, en vérifiant cependant que la correction demeure faible. Avec la procédure décrite ci-dessus, le modèle choisi in fine pour imputer est le même modèle que celui construit en section 3.

Finalement, il est important d'observer qu'une imputation des données provisoires manquantes ou incomplètes à l'aide d'un modèle qui se base seulement sur des données provisoires présente un inconvénient important. La valeur imputée sur la base d'un tel modèle ne tient pas compte des données définitives qui sont pourtant disponibles. Cette approche présente le risque d'imputer une valeur provisoire plus élevée que la valeur définitive, ce qui est par définition un non-sens. Pour prendre en compte l'information apportée par les données définitives, la solution retenue dans tout ce travail est d'effectuer l'imputation sur la différence $\text{diff}_{E,t}$ entre les données définitives et les données provisoires, chaque fois que la valeur définitive est disponible. À partir d'une telle imputation et de la valeur définitive, il est facile de déduire une imputation pour les données provisoires, et cette imputation sera forcément inférieure aux données définitives tant que la valeur imputée pour $\text{diff}_{E,t}$ est positive. Pour conclure, il faut mentionner que les imputations concernent un nombre limité d'observations. Leur influence sur les résultats finaux est étudiée et demeure faible.

3 Modélisation

3.1 Présentation générale des modèles REG-ARIMA

L'outil X13-ARIMA-SEATS utilisé dans ce travail à travers le package R *seasonal* permet d'ajuster des modèles REG-ARIMA aux données qui sont ensuite utilisés pour l'interprétation et pour obtenir des prédictions. Ces prédictions servent de base in fine pour le calcul des flash estimates. Une description complète des modèles REG-ARIMA est donnée dans U.S. Census Bureau [2025], section 4.1. Comme son nom l'indique, les modèles REG-ARIMA combinent la régression linéaire classique, qui cherche à prédire linéairement une réponse à partir de prédicteurs bien déterminés, et les modèles ARIMA développés par exemple dans Brockwell & Davis [2009]. L'équation générale du modèle pour une série de données y_t est :

$$\phi(B)\Phi(B^s)(1-B)^d(1-B^s)^D\left(y_t - \sum_{i=1}^{n_{\text{pred}}} \beta_i x_{i,t}\right) = \theta(B)\Theta(B^s)\varepsilon_t \tag{2}$$

où

- d et D sont respectivement les degrés de différenciation et différenciation saisonnière. Dans tout ce travail, d et D sont toujours fixés à 1.
- s est la période. Dans tout ce travail, les données sont mensuelles et ainsi $s = 12$.
- B est l'opérateur de retard (lag) qui transforme une série y_t en $By_t := y_{t-1}$.
- $\phi(\cdot)$ et $\Phi(\cdot)$ sont deux polynômes de degrés respectifs $p \geq 0$ et $P \geq 0$ associés à la partie autorégressive (AR) du modèle ARIMA saisonnier : $\phi(B) = 1 - \phi_1 B - \ldots - \phi_p B^p$ et $\Phi(B^s) = 1 - \Phi_1 B^s - \ldots - \Phi_P B^{sP}$. Les nombres $\phi_1, \ldots, \phi_p$ et $\Phi_1, \ldots, \Phi_P$ sont des paramètres du modèle.
- $\theta(\cdot)$ et $\Theta(\cdot)$ sont deux polynômes de degrés respectifs $q \geq 0$ et $Q \geq 0$ associés à la partie moyenne mobile (MA) du modèle ARIMA saisonnier : $\theta(B) = 1 + \theta_1 B + \ldots + \theta_q B^q$ et $\Theta(B^s) = 1 + \Theta_1 B^s + \ldots + \Theta_Q B^{sQ}$. Les nombres $\theta_1, \ldots, \theta_q$ et $\Theta_1, \ldots, \Theta_Q$ sont des paramètres du modèle.
- n_{pred} est le nombre de prédicteurs dans la partie REG du modèle REG-ARIMA. Cela inclue les prédicteurs pour modéliser les effets calendrier et les prédicteurs associés aux outliers.
- $x_{1,t}, \ldots, x_{n_{\text{pred}},t}$ sont n_{pred} séries temporelles déterministes qui servent de prédicteurs pour y_t dans la partie REG du modèle REG-ARIMA, et $\beta_1, \ldots, \beta_{n_{\text{pred}}}$ sont les coefficients de régression associés.
- ε_t est un bruit blanc de variance $\sigma^2 > 0$.

L'équation (2) est l'équation classique d'un modèle ARIMA saisonnier appliquée à une série $y_t - \sum_{i=1}^{n_{\text{pred}}} \beta_i x_{i,t}$ dont les outliers et les effets calendrier ont déjà été filtrés par la partie REG du modèle. Les modèles ARIMA saisonniers sont présentés dans Brockwell & Davis [2009] et sont des modèles classiques d'analyse de séries temporelles. Les paramètres d'un modèle REG-ARIMA sont $\phi_1, \ldots, \phi_p, \Phi_1, \ldots, \Phi_P, \theta_1, \ldots, \theta_q, \Theta_1, \ldots, \Theta_Q, \sigma^2, \beta_1, \ldots, \beta_{n_{\text{pred}}}$ et sont tous des nombres réels. Les nombres p, P, q, Q sont des entiers positifs ou nuls qui doivent être estimés. Les prédicteurs $x_{i,t}, 1 \leq i \leq n_{\text{pred}}$ dans la partie REG doivent aussi être déterminés, tout comme le nombre de prédicteurs n_{pred}. Au total, le modèle comporte $p+P+q+Q+n_{\text{pred}}+1$ paramètres, ce qui peut être une valeur importante, voir section 3.7. Dans ce travail, la valeur de la somme $p + P + q + Q$ ne dépasse pas 4. En revanche, la valeur de n_{pred} peut dépasser 15 selon le cas, si bien que le nombre total de paramètres est parfois de l'ordre de 20, et plus souvent autour de 10.

Lorsque les entiers $p, P, q, Q, n_{\text{pred}}$ et les séries temporelles $x_{i,t}$ sont fixés, les paramètres du modèle REG-ARIMA sont estimés à l'aide de X13-ARIMA-SEATS par une procédure itérative qui

estime successivement les paramètres du modèle de régression puis les paramètres du modèle ARIMA saisonnier. Ces procédures sont répétées jusqu'à convergence, ce qui est décrit dans U.S. Census Bureau [2025] en section 4.5.

Finalement, il convient de dire que les modèles impliqués font une hypothèse de stationnarité. Dans le modèle REG-ARIMA de l'équation (2), la série

$$(1-B)^d(1-B^s)^D\left(y_t - \sum_{i=1}^{n_{\text{pred}}} \beta_i x_{i,t}\right)$$

est stationnaire. En d'autres termes la série initiale corrigée de ses outliers et effets calendrier, puis différenciée, présente une stationnarité modélisée à l'aide de modèles ARMA (et non ARIMA) saisonnier. Ce point justifie l'importance d'une modélisation appropriée pour les effets calendrier et outliers. La présence d'outliers résiduels, qui ne seraient pas traités et corrigés, viendrait en effet rompre l'hypothèse de stationnarité qui est utilisée dans toute la modélisation.

3.2 Choisir un modèle REG-ARIMA

Construire un modèle REG-ARIMA consiste dans ce travail à faire plusieurs choix. Pour la partie REG du modèle, des outliers sont détectés et traités selon une procédure décrite en section 3.2.1. La partie REG implique aussi un choix de modélisation quant aux effets calendrier. Ce travail est décrit en section 3.2.2. Pour la partie ARIMA du modèle REG-ARIMA, la procédure de construction du modèle est décrite en section 3.2.3. Finalement, un résumé de la méthodologie de choix du modèle appliquée pour les MDND est donné en section 3.2.4

3.2.1 Transformation préliminaire et outliers

X13-ARIMA-SEATS donne la possibilité d'ajuster les modèles sur les données avec ou sans une transformation logarithmique (log), c'est-à-dire en remplaçant ou non les données par leur logarithme. Les analyses sur ce point ne sont pas présentées dans ce rapport mais pointent toutes en faveur d'une transformation log. Ainsi, pour tous les modèles choisis dans ce rapport, une transformation en log est effectuée, ce qui signifie que tous les modèles présentés s'appliquent au logarithme des données et non aux données elles-mêmes. Un argument essentiel en faveur de la transformation log est que les données prédites prennent nécessairement des valeurs positives et ne prennent jamais de valeurs nulles compte tenu de la taille du domaine étudié : toute la Suisse.

Pour la partie REG du modèle REG-ARIMA, X13-ARIMA-SEATS offre la possibilité de gérer les outliers , tels que définis dans ce rapport en section 2.3. Les choix offerts aux utilisateurs sont détaillés dans U.S. Census Bureau [2025], section 4.3. Dans ce travail, les outliers choisis sont toujours des "Additive Outliers" (AO) ou bien des "Level Shifts" (LS). Les abréviations AO et LS seront utilisées un grand nombre de fois dans ce rapport.

Les AO correspondent à une observation unique, pour un mois d'une année donnée, qui est considérée comme soit très haute soit très basse. Dans l'équation (2), un AO à une date $t^\star$, par exemple janvier 2020, est représenté en utilisant une série temporelle $x_{i,t}$ égale à 1 à la date $t^\star$ et 0 partout ailleurs. Le coefficient β_i associé vient alors corriger l'observation $y_{t^\star}$ en $y_{t^\star} - \beta_i$, et laisse le reste de la série y inchangée. Le signe de β_i informe sur la nature de l'observation $y_{t^\star}$, c'est-à-dire sur le fait que l'observation est considérée comme très élevée ($\beta_i > 0$) ou très faible ($\beta_i < 0$). La valeur absolue de β_i informe sur l'amplitude de la correction appliquée à y_t. Un autre

indicateur qui informe sur l'amplitude de l'outlier est la t-valeur associée à β_i, c'est-à-dire le ratio entre β_i et son écart-type estimé. Les outliers sont précisément détectés lorsque cette t-valeur dépasse un seuil, voir les explications en fin de cette section.

Les LS sont utilisés pour prendre en compte un changement de niveau durable de la série temporelle, c'est-à-dire un saut soit vers le haut soit vers le bas qui persiste dans le temps. Dans l'équation (2), un LS à une date t se modélise avec une série temporelle $x_{i,t}$ égale à -1 avant la date t et 0 à partir de, et en incluant, la date t, voir U.S. Census Bureau [2025] section 4.3.

Une détection automatique d'outliers est proposée par X13-ARIMA-SEATS et celle-ci est utilisée pour construire un premier jeu de potentiels outliers. Une fois la détection automatique effectuée, une analyse approfondie et une discussion avec la section DEM sont effectuées pour valider, ou non, les outliers détectés automatiquement. Pour chaque observation effectivement validée comme outlier, AO ou LS, une interprétation discutée avec la section DEM est tirée. L'objectif lors de la production des flash estimates est de conserver le même jeu d'outliers dans le temps et de ne plus utiliser la procédure de détection automatique. Une utilisation systématique de cette procédure automatique présente en effet le risque d'obtenir des jeux d'outliers différents selon les années de production des flash estimates et ainsi des modèles potentiellement très différents. Les choix d'outliers pour chaque série sont détaillés en sections 3.3.2, 3.4.2, 3.5.2 et 3.6.2

Un paramètre important de la procédure de détection automatique d'outlier proposée par X13-ARIMA-SEATS est un seuil nommé "outlier.critical". Ce paramètre est, comme évoqué précédemment, un seuil limite sur la t-valeur associée au coefficient β_i d'un outlier donné. Le nombre d'outliers détectés automatiquement tend à augmenter lorsque le seuil est bas, et inversement. Les détails sur la procédure automatique de détection et sur le seuil sont donnés dans U.S. Census Bureau [2025], en section 7.11. La valeur par défaut du seuil dépend du nombre d'observations. Pour des séries avec 14 années de données (168 observations), la valeur par défaut du seuil est fixée à 3.92.

3.2.2 Effets calendrier

Le second jeu de prédicteurs présents dans la partie REG du modèle REG-ARIMA sont les prédicteurs associés aux effets calendrier. La plupart des séries temporelles mensuelles sont effectivement influencées par la structure du calendrier, c'est-à-dire par le nombre de lundis, mardis, ..., dimanches et jours fériés au cours du mois. Le travail à effectuer ici consiste à déterminer si les différents jours non-fériés de la semaine (lundis, mardis, ..., samedis non fériés) ont une influence similaire ou non et si les jours fériés ont une influence identique ou non à celle d'un dimanche.

Pour effectuer cette tâche, la section METH se base sur un calendrier Suisse des jours fériés et sur un rapport de méthodes de l'OFS dédié à cette question, voir Chevalier [2023]. Pour chacun des 8 modèles évalués dans ce travail, plusieurs jeux de prédicteurs associés aux effets calendrier sont testés. Chaque jeu de prédicteurs fait des hypothèses différentes sur l'influence des jours de la semaine et des jours fériés. Lorsque certains jours ont une influence identique, le nombre de prédicteurs est réduit. S'il est supposé que tous les jours de la semaine et/ou les jours fériés ont une influence différente, le nombre de prédicteurs est plus grand. L'objectif ici est de bien capter les effets calendrier dans le modèle et ce avec le modèle le plus parcimonieux possible, c'est-à-dire contenant un nombre de prédicteurs le plus faible possible. Les choix concernant les effets calendrier pour chaque série sont détaillés en sections 3.3.3, 3.4.3,

3.5.3 et 3.6.3. Les outils utilisés pour ces choix, notamment les différents jeux de prédicteurs considérés, sont détaillés dans Chevalier [2023].

Pour traiter au mieux les effets calendrier, ce sont des groupes de prédicteurs qui seront comparés entre eux, correspondant chacun à une "option". Voici un rapide résumé des options possibles, détaillées dans Chevalier [2023] :

1. Option 1 : dans cette variante, la plus complexe, les lundis, mardis, ..., samedis non-fériés ainsi que les lundis, mardis, ..., samedis fériés ont tous un effet différent. Les dimanches non-fériés et dimanches fériés ont le même effet. Ce modèle contient 13 paramètres. L'option 1 n'est jamais choisie dans ce travail en raison du nombre de paramètres trop important.
2. Option 2 : cette variante vient simplifier l'option 1, en donnant un effet identique à tous les jours fériés. Les lundis, mardis, ..., samedis non-fériés ont toujours un effet différent. Ce modèle contient 7 paramètres. L'option 2 est choisie pour la moitié des 8 modèles étudiés ici.
3. Option 3 : les lundis, mardis, jusqu'aux vendredis non-fériés ont le même effet. Les samedis non-fériés ont leur propre effet, et les dimanches et jours fériés ont le même effet, comme dans l'option 2. Ce modèle contient 3 paramètres. L'option 3 est choisie pour la moitié des 8 modèles étudiés.
4. Option 4 : les lundis, mardis, jusqu'aux vendredis non-fériés ont le même effet. Les samedis, dimanches et jours fériés ont tous le même effet. Le modèle n'a plus que 2 paramètres et n'est jamais choisi ici.

Dans ce travail, les séries ont au maximum 168 observations, soit 14 ans de données mensuelles. Compte tenu du nombre souvent important de paramètres dédiés à la gestion des outliers et au modèle ARIMA saisonnier, l'option 1 ci-dessus n'est jamais choisie. Un traitement différencié des jours fériés pourrait dans certains cas être intéressant pour gérer par exemple des effets de ponts ou week-end prolongés, souvent propices aux mariages par exemple. Cependant, un tel traitement devrait se faire dans le cadre d'un modèle ad-hoc, avec moins de 13 paramètres. Ceci n'a pas été effectué ici et constitue une perspective d'amélioration possible.

Pour choisir entre les options 2, 3 ou 4, la procédure décrite dans Chevalier [2023], section 4, est utilisée. En particulier, la série dite "corrigée de ses effets calendrier", calculée par X13-ARIMA-SEATS doit effectivement ne plus avoir d'effets calendrier, ce qui peut se vérifier à l'aide d'une analyse spectrale, voir Chevalier [2023], sections 4.4 et 4.5. Dans ce rapport, la série corrigée de ses effets calendrier est aussi appelée "série désaisonnalisée". Il s'agit d'une série dont la saisonnalité a été estimée puis filtrée par X13-ARIMA-SEATS, et dont les effets calendrier devraient avoir été supprimés, comme expliqué. Cette série, calculée par X13-ARIMA-SEATS, n'est pas utilisée pour les prédictions et sert simplement à choisir l'option la plus adaptée.

L'option choisie in fine est ainsi l'option la plus parcimonieuse en nombre de paramètres qui permet effectivement de corriger les effets calendrier au vu des analyses spectrales.

3.2.3 Choix du modèle ARIMA saisonnier

Une fois les prédicteurs dans la partie REG choisis, X13-ARIMA-SEATS est capable d'estimer tous les coefficients β_i associés présents dans l'équation (2), mais aussi de proposer un modèle ARIMA saisonnier, c'est-à-dire de suggérer des valeurs pour les entiers, p, P, q, Q et d'estimer les $p+P+q+Q+1$ paramètres associés à la partie ARIMA du modèle. Le choix des entiers p, P, q et

Q se fait sur la base de critères d'informations du type "Akaike corrigé" et est détaillé dans U.S. Census Bureau [2025] en section 5.5. Le choix automatique proposé par X13-ARIMA-SEATS est utilisé pour avoir des valeurs initiales pour les entiers p, P, q et Q. Selon le cas, ces valeurs sont modifiées par la suite ou non. En particulier, une analyse des autocorrélations des résidus du modèle est effectuée pour valider définitivement le nombre de paramètres du modèle ARIMA. Les détails pour chaque série sont donnés en sections 3.3.4, 3.4.4, 3.5.4 et 3.6.4.

Pour assurer une stabilité dans le temps du modèle ARIMA, le choix des ordres p, P, q et Q ne change pas lorsque les données pour une nouvelle année sont disponibles. De cette manière, tout comme pour les paramètres associés à la partie REG, le nombre de paramètres est fixé de sorte que les modèles utilisés pour produire les flash estimates restent relativement similaires d'une année à l'autre. Le nombre de paramètres du modèle ne vient changer que si un nouvel outlier est détecté dans les données les plus récentes. Les valeurs des paramètres du modèle ARIMA saisonnier (i.e. $\phi_1, \ldots, \phi_p, \theta_1, \ldots, \theta_q$, etc.) sont par contre systématiquement ré-estimées lorsque de nouvelles données sont disponibles, tout comme les valeurs des n_{pred} coefficients β_i.

X13-ARIMA-SEATS permet normalement un choix automatique des degrés de différentiation d et D du modèle ARIMA(p,d,q)(P,D,Q) [3]. Après analyse, il a été décidé de toujours fixer ces degrés à 1. Ceci est une approche relativement standard puisqu'il faut rappeler que les modèles ARIMA saisonniers font l'hypothèse que la série différenciée est stationnaire. Dans cette optique, la différenciation à l'ordre 1 (d=1) sert à supprimer la tendance qui existe sur les données et la différenciation saisonnière d'ordre 1 (D=1) sert à supprimer la saisonnalité. L'autre avantage de ce choix est de faciliter la comparaison des modèles entre eux : comparer des modèles dont les degrés de différenciation sont différents est en effet une tâche délicate.

Si le choix de toujours procéder à une différenciation (D=1) paraît peu contestable au vue de la saisonnalité claire de toutes les séries étudiées, le choix de toujours appliquer une différenciation d'ordre 1 (d=1) est plus discutable dans la mesure où la tendance de certaines séries peut être vue comme faible. Appliquer une différenciation d'ordre 1 à une série dont la tendance est faible n'est pas anodin et peut conduire à un des effets de tendance dans les prédictions du modèle, ce qui n'est pas forcément souhaitable. Ce problème est discuté pour les mariages et les divorces, dans le cas de la différence entre les données définitives et provisoires, voir sections 3.3.4 et 3.4.4.

3.2.4 Vue synthétique de la procédure du choix de modèle

Voici un résumé des différentes étapes du choix de modèle, appliquées pour les MDND dans les sections suivantes.

1. Imputations. Les données provisoires manquantes ou incomplètes doivent être imputées, et cette imputation se fait sur la base d'un modèle sur la série $\text{diff}_{E,t}$. Chaque imputation se fait en deux temps. Une première imputation est calculée en laissant globalement plus de latitude à X13-ARIMA-SEATS dans le choix du modèle. Dans cette première étape, l'option 3 est choisie pour modéliser les effets calendrier, sauf pour les mariages où c'est l'option 2 qui est sélectionnée. X13-ARIMA-SEATS choisit librement les outliers de la série, ainsi que le modèle ARIMA. Il est aussi rappelé qu'une transformation en log est toujours choisie et que les d et D du modèle ARIMA saisonnier sont toujours fixés à 1. Une fois cette première étape terminée, une imputation est disponible, ce qui permet d'avancer dans les

3. Dans ce rapport, les modèles ARIMA saisonniers sont notés ARIMA(p,d,q)(P,D,Q). Une autre notation fréquente précise la période s de la série, ce qui donne ARIMA(p,d,q)(P,D,Q)$_s$. Dans ce travail, la période s vaut toujours 12.

étapes 2, 3 et 4 énumérées ci-dessous. Lorsqu'un modèle final est choisi pour la série $\text{diff}_{E,t}$ à l'issue de l'étape 4, celui-ci est utilisé pour réviser l'imputation obtenue à cette première étape. Il est alors vérifié si la nouvelle imputation est proche de la précédente. Les étapes 2, 3 et 4 décrites ci-dessous sont ré-évaluées avec cette nouvelle imputation.

2. Outliers. Des outliers sont détectés automatiquement avec X13-ARIMA-SEATS, pour les séries $\text{prov}_{E,t}$ et $\text{diff}_{E,t}$. L'option utilisée pour modéliser les effets calendrier ainsi que le choix du modèle ARIMA sont susceptibles d'avoir une influence sur la détection d'outliers. Si le modèle ARIMA choisi automatiquement en même temps que la détection d'outliers diffère du modèle ARIMA choisi in fine il est important de vérifier si le choix final de modèle ARIMA a une influence sur la détection automatique d'outliers. Chacun des outliers trouvés par détection automatique doit avoir une interprétation validée avec la section DEM. Il peut arriver que certains outliers soient détectés avec une t-valeur qui dépasse de peu le seuil de détection. Ceci n'est pas gênant si des interprétations crédibles existent pour justifier l'outlier. Dans le cas contraire, le seuil de détection peut être relevé. Il s'avère que ceci est appliqué au moins une fois pour les 4 séries étudiées. Des justifications sont données sur ces choix.
3. Effets calendrier. Le choix relatif aux effets calendrier revient à décider quel jeu de prédicteurs est sélectionné pour modéliser ces effets. Les jeux de prédicteurs envisagés correspondent aux options 2, 3 ou 4 décrites en section 3.2.2. L'objectif est de choisir l'option la plus parcimonieuse en termes de nombre de paramètres qui donne de bons résultats sur les analyses spectrales des résidus du modèle, c'est-à-dire qui indiquent que les résidus ne présentent plus d'effets calendrier. Dans un premier temps, la recherche de la meilleure option se fait en laissant X13-ARIMA-SEATS choisir automatiquement le modèle ARIMA de l'étape 4 suivante. Si une certaine option est retenue et si le modèle ARIMA finalement choisi diffère de celui sélectionné automatiquement, il convient de vérifier que les analyses spectrales demeurent satisfaisantes avec le modèle final.
4. Modèle ARIMA. Une fois que toute la partie REG du modèle REG-ARIMA est déterminée, cette étape consiste à examiner la proposition automatique de modèle ARIMA donnée par X13-ARIMA-SEATS. Un modèle est accepté ou non sur des critères classiques décrits dans Brockwell & Davis [2009] et mentionnés ci-après. D'une part, les paramètres du modèle ARIMA doivent être significativement différents de zéro, ce qui peut être évalué à l'aide de p-valeurs associées à chaque paramètre. Ceci assure que le modèle ne contient pas de paramètres inutiles. D'autre part, les résidus du modèle ne doivent pas présenter d'autocorrélations fortes, ce qui est évalué à l'aide de tests de type Ljung-Box. Une autocorrélation significative dans les résidus indiquerait en effet que la corrélation dans les données n'est pas pleinement captée par le modèle. Dans les sections suivantes, par exemple pour les mariages ou les naissances, il arrive que la proposition faite automatiquement par X13-ARIMA-SEATS ne soit pas jugée satisfaisante et qu'un autre modèle ARIMA soit choisi in fine.

3.3 Modélisation des mariages

3.3.1 Imputations

La figure 1 présente les données des mariages livrées par la section DEM. Les données provisoires sont indiquées sur le graphique de gauche. La différence entre les données définitives et les données provisoires est représentée sur le graphique de droite. Pour les mariages, comme indiqué précédemment, les données de décembre 2011, 2012 et 2013 sont manquantes. Aucune valeur ne semble être incomplète et le travail d'imputation consiste seulement à imputer les valeurs manquantes.

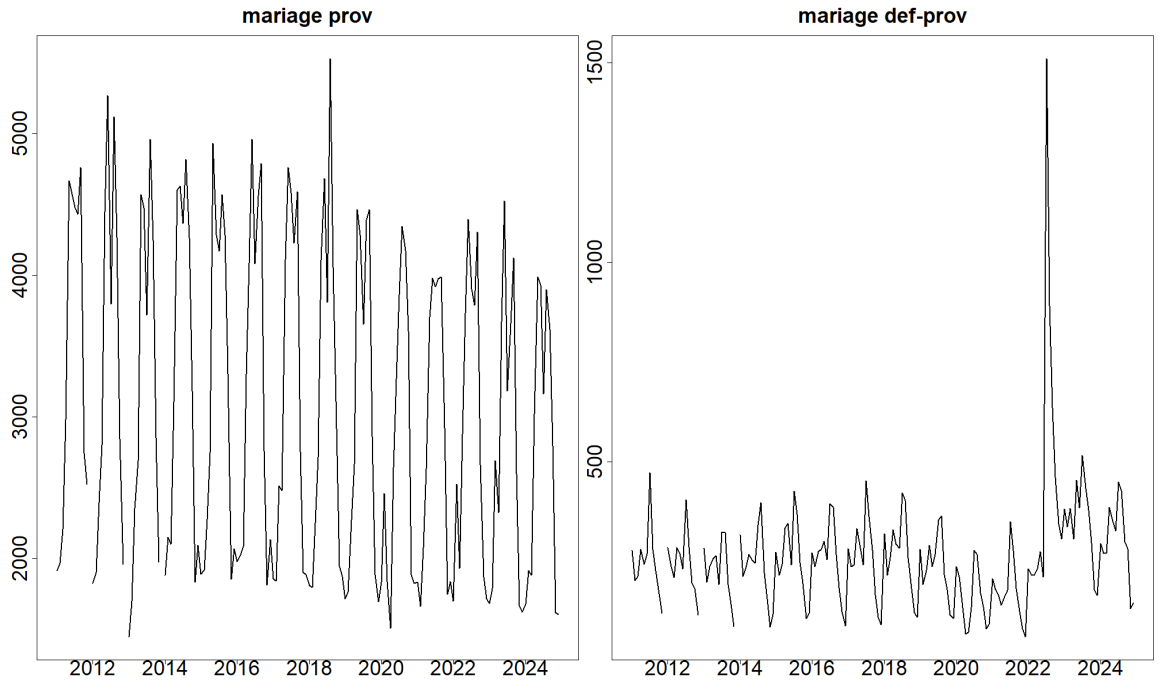


FIGURE 1 Nombre mensuel de mariages en Suisse : données provisoires (à gauche) et différence entre les données définitives et les données provisoires (à droite).

Comme expliqué en section 2.3, une première imputation est effectuée avec un modèle qui n'est pas toujours le modèle final. Pour imputer, les données manquantes de $\text{diff}_{E,t}$ sont remplacées par la valeur arbitraire de 9'999 de sorte à ce que les observations de décembre 2011, 2012 et 2013 soient traitées par X13-ARIMA-SEATS comme des additive outliers, AO. La première imputation est faite avec le traitement des effets calendrier selon l'option 2, ce qui est l'option choisie dans le modèle final.

X13-ARIMA-SEATS choisit dans ce cas un modèle assez similaire au modèle qui sera choisi in fine, avec des différences légères sur le choix des outliers. En utilisant ce modèle et la composante outlier de l'équation (1), les premières valeurs imputées pour $\text{diff}_{E,t}$ pour décembre 2011, 2012 et 2013 sont respectivement de 102, 98 et 103. Si c'est le modèle final, inconnu au départ, qui est utilisé pour recalculer les imputations, les valeurs imputées sont respectivement de 105, 101 et 101. Ces nouvelles imputations sont très proches des valeurs précédentes compte tenu de l'ordre de grandeur du nombre de mariages en un mois en Suisse, qui va de 2000 à 5000 environ. La sensibilité des imputations à la valeur arbitraire de 9'999 choisie pour remplacer les valeurs manquantes a été testée et reste nulle du moment que cette valeur est suffisamment éloignée des valeurs plausibles prises par la série $\text{diff}_{E,t}$.

Les données de la figure 1 complétées par les imputations sont représentées en figure 2. Si les imputations effectuées sur la série $\text{diff}_{E,t}$ (voir figure 2, graphique de droite) peuvent sembler assez naturelles avec des imputations autour de 100 pour le mois de décembre de trois années consécutives, l'imputation correspondante pour les données provisoires (figure 2, graphique de gauche), obtenue à l'aide de $\text{diff}_{E,t}$ et des données définitives, est beaucoup moins intuitive. Par exemple, l'imputation de décembre 2012 pour les données provisoires est une valeur sensiblement plus haute que les deux valeurs adjacentes de novembre 2012 et janvier 2013. Il s'avère en effet que le nombre de mariages en décembre 2012 a été particulièrement important du fait d'un effet de date insolite : un nombre important de couples a souhaité se marier la semaine

du 12 décembre 2012 : 12.12.12. Ceci illustre l'intérêt d'imputer les données provisoires à l'aide d'un modèle sur $\text{diff}_{E,t}$, puisqu'une telle imputation prend en compte l'information des données définitives.

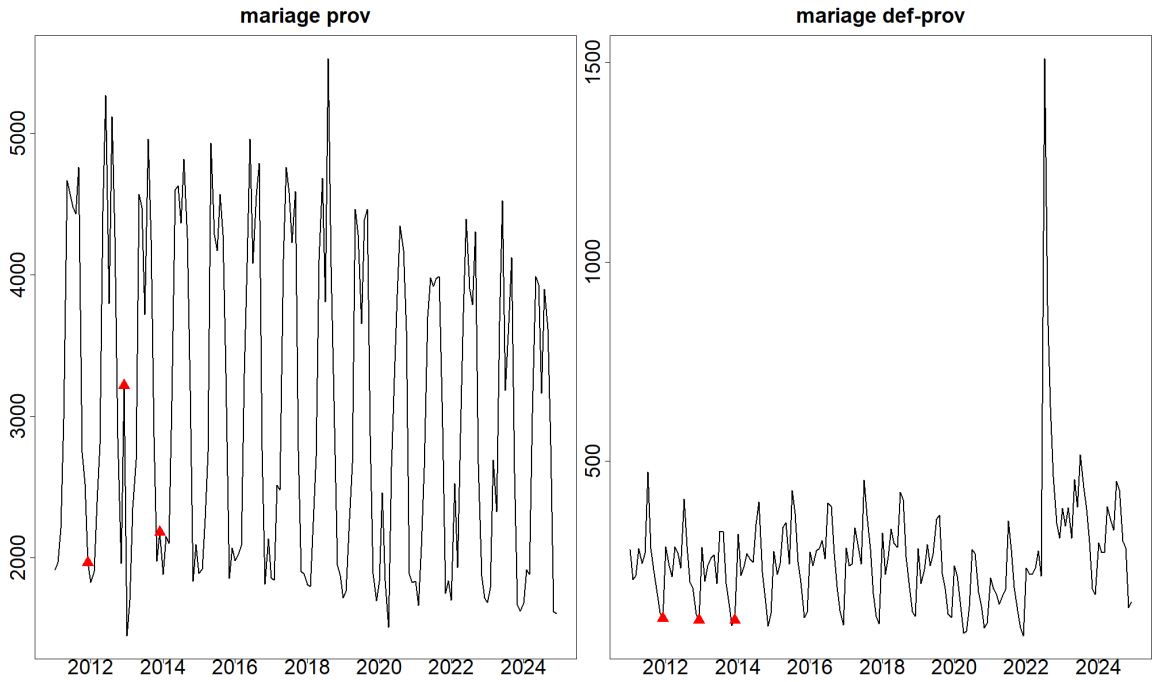


FIGURE 2 Nombre mensuel de mariages en Suisse : données provisoires à gauche et différence entre les données définitives et les données provisoires à droite. Les données imputées à l'aide du modèle final sur $\text{diff}_{E,t}$ sont indiquées par les triangles rouges.

3.3.2 Outliers

Parmi les 4 types de séries MDND étudiés dans ce rapport les mariages sont peut-être les plus complexes à étudier compte tenu de la difficulté à bien modéliser les outliers.

Pour les données provisoires, la procédure automatique de détection d'outliers appliquée en utilisant l'option 2 pour les effets calendrier, détecte au total 10 outliers, avec le seuil de détection par défaut. Ces outliers sont tous des "additive outliers", AO. Tous ces outliers ont été discutés avec la section DEM qui a chaque fois validé une interprétation crédible. La liste des 10 outliers avec leur interprétation est donnée ci-dessous :

1. AO en novembre 2011. Le nombre de mariages élevé de ce mois-ci peut s'expliquer par un effet de date insolite. Le 11 novembre 2011 (11.11.11) tombe un vendredi.
2. AO en décembre 2012. Le nombre de mariages particulièrement élevé s'explique toujours par un effet de date insolite. Il faut noter ici que l'observation qui est considérée comme un AO est une observation imputée. Le fait que cette observation soit un outlier ne remet cependant pas en question la technique d'imputation, puisque les données définitives montrent clairement une valeur très élevée en décembre 2012.
3. AO en janvier 2013. Ici le nombre de mariages est particulièrement bas par rapport à un mois de janvier d'une autre année. Ceci peut s'interpréter comme un effet de compensation par rapport au nombre très élevé de mariages le mois précédent. En résumé : les couples qui souhaitaient se marier lors de cette période de décembre/janvier ont préféré le mois de décembre 2012, peut-être en raison de la date insolite, plutôt que janvier 2013.

4. AO en février 2020. Le nombre de mariages est élevé et s'interprète encore en raison de dates insolites / dates palindromes. Le 2 février 2020 (02.02.2020) ou bien le 20 février 2020 (20.2.20) sont des exemples.
5. AO en avril 2020. En raison du confinement lié au début de la pandémie de COVID-19, le nombre de mariages est faible ce mois-ci. Beaucoup de mariages prévus ont probablement été reportés.
6. AO en mai 2020. Le nombre de mariages est également plus faible que prévu, mais dans des proportions moindres que le mois précédent d'après le coefficient β_i associé à cet AO. L'interprétation est assez aisée, avec un prolongement du confinement.
7. AO en juin 2020. L'interprétation ne change pas. Il faut noter ici aussi que le coefficient β_i associé à cet outlier est moins fort, c'est-à-dire moins négatif, que celui de mai 2020, qui était lui-même nettement moins fort que celui d'avril 2020. Ceci traduit une normalisation progressive de la situation en termes de nombre de mariages. Il n'y a d'ailleurs plus d'outliers détectés en juillet ou août 2020, mais il n'y a pas non plus une explosion à la hausse des mariages cet été là. Cela montre que les mariages annulés en avril, mai et juin 2020 n'ont pas été systématiquement célébrés durant l'été 2020. Il faut garder à l'esprit qu'au début du mois de juin 2020, il n'y avait pas encore de fin annoncée du confinement. Dans ces conditions, il n'était peut-être pas facile de planifier une fête de mariage en juillet ou août. Ceci explique peut-être l'absence d'outlier "positif" (c'est-à-dire associés à un coefficient β_i positif) en juillet, août.
8. AO en octobre 2020. Le nombre de mariages est plus élevé que prévu sur ce mois qui est habituellement assez faible en nombre de mariages. Ceci s'interprète à l'aide des reports évoqués précédemment. Au cours du mois d'octobre 2020, le confinement est levé depuis environ 4 mois. En Suisse, les mesures liées à la seconde vague du COVID-19 ne débutent qu'à la fin du mois d'octobre 2020 et sont au départ moins restrictives (en termes de fermetures de commerces, etc.) que lors de la première vague.
9. AO en février 2022. Le coefficient β_i associé à cet AO est positif et très élevé, ce qui traduit un nombre de mariages beaucoup plus élevé que la valeur attendue. Ceci s'interprète par un effet de dates insolites : durant ce mois il y a le 2 février 2022 (2.2.22) ou bien le 22 février 2022 (22.2.22).
10. AO en mars 2023. Le coefficient β_i associé à cet AO est positif, mais moins fort que celui de février 2022. Cet outlier peut s'expliquer encore une fois par un effet de date insolite, avec le 23 mars 2023 (23.3.23).

Il est constaté que les outliers pour les mariages sont souvent liés à des effets de date insolite. Dans cette optique, il est légitime de s'attendre à un nombre de mariages assez élevé en avril 2024 (24.4.24), en mai 2025 (25.5.25) ou bien en juin 2026 (26.6.26). À l'heure où ce rapport est écrit, les valeurs des mois d'avril 2024 et de mai 2025 sont effectivement relativement élevées, mais pas suffisamment pour être considérées comme des outliers par X13-ARIMA-SEATS. Pour éviter d'ajouter un paramètre au modèle pour chaque mois contenant une date insolite, une idée pourrait être de travailler avec un régresseur $x_{i,t}$ spécifique aux dates insolites de sorte à traiter plusieurs, voire toutes les dates insolites simultanément avec un nombre réduit de prédicteurs. Ceci est évoqué en section 3.3.4 mais n'a pas été développé pour le moment.

En plus des données provisoires, un modèle est évalué pour $\text{diff}_{E,t}$, et ce modèle comprend ses propres outliers. La figure 2, graphique de droite, montre que la série se comporte de manière relativement régulière, avec notamment une saisonnalité claire, jusqu'au milieu de l'année 2022. À partir de là, la série prend des valeurs plus élevées que par le passé, avec un pic très élevé à l'été 2022. Sur l'année 2024, la série semble lentement revenir à la situation antérieure, mais

tout de même avec des valeurs globalement plus élevées. Le pic important observé à l'été 2022 s'explique par la reconnaissance en Suisse du mariage de couples du même sexe, en vigueur depuis le 1er juillet 2022. La reconnaissance fait suite à une acceptation par la population en votation en septembre 2021. Un nombre important de couples de personnes de même sexe qui souhaitaient se marier l'ont fait dès le mois de juillet 2022, ce qui explique ce pic important. Par ailleurs, il faut noter que les mariages de personnes du même sexe ne sont pas enregistrés dans les données provisoires, y compris après juillet 2022. Dès lors, comme ceux-ci figurent bien dans les données définitives, la différence $\text{diff}_{E,t}$ change de comportement à partir de cette date.

Le fait de ne pas enregistrer ces mariages dans les données provisoires complique la prédiction de $\text{diff}_{E,t}$, sans toutefois rendre celle-ci impossible, du moins à long terme. Dans un modèle REG-ARIMA, le changement de comportement peut se modéliser à l'aide de level shifts, LS. Par ailleurs, la série des mariages semble revenir progressivement à une phase plus stationnaire, avec une saisonnalité qui devient relativement régulière. Une telle série est donc difficile à prédire en 2023 ou 2024, mais devient plus facile à prédire pour les années suivantes. Pour tempérer l'optimisme, il convient néanmoins de dire qu'en comparaison avec les trois autres évènements étudiés ici (divorces, naissances, décès) les performances des modèles de prédiction des mariages demeurent moins bonnes, comme le montre la section 4.2.

Fort de ces remarques, les 4 outliers suivants sont retenus pour la série $\text{diff}_{E,t}$. Tous ces outliers sont des Level Shifts (LS) :

1. LS en avril 2020. Un coefficient β_i négatif montre une chute durable du nombre de mariages à partir d'avril 2020. La date d'avril 2020 était déjà un outlier pour les données provisoires, avec un faible nombre de mariages en avril, mai et juin 2020 qui était traité avec trois AO. La chute s'observe donc aussi sur la série $\text{diff}_{E,t}$, ce qui n'est pas surprenant. L'interprétation est inchangée.
2. LS en juin 2020. Le coefficient β_i associé est positif, mais un peu moins élevé en valeur absolue que le coefficient d'avril 2020. Ceci traduit que $\text{diff}_{E,t}$ se normalise un peu à partir de juin 2020, mais pas totalement car le coefficient β_i associé est plus bas en valeur absolue que celui d'avril 2020. Après juin 2020, le niveau de $\text{diff}_{E,t}$ demeure donc plus bas que le niveau d'avant avril 2020.
3. LS en juillet 2022. Le coefficient β_i associé est positif et très grand, ce qui est lié à la reconnaissance des mariages de personnes du même sexe, voir les explications précédentes.
4. LS en janvier 2023. Le coefficient β_i associé est négatif, mais moins élevé en valeur absolue que celui de juillet 2022. Ceci traduit un retour progressif à une saisonnalité régulière, avec la fin progressive du nombre très élevé de mariages.

Les outliers donnés ici sont obtenus en relevant légèrement le seuil "outlier.critical" à 4.5, ce qui est une valeur plus élevée que la valeur par défaut de 3.92. Le seuil par défaut donnait des outliers supplémentaires plus difficiles à interpréter et n'a donc pas été retenu.

La figure 3 représente la composante multiplicative "outliers" obtenue avec ces quatre LS, ainsi que les séries avant (graphique de droite, courbe noire) et après (courbe rouge pontillée) correction des outliers. C'est sur cette série corrigée des outliers que la suite du travail est effectuée, notamment la détection des effets calendrier, le choix du modèle ARIMA saisonnier, etc. Il est possible de voir que la composante outlier calculée permet de lisser partiellement, mais pas totalement, le pic de juillet 2022. Les données avant cette date sont relevées légèrement pour mieux coller aux données récentes.

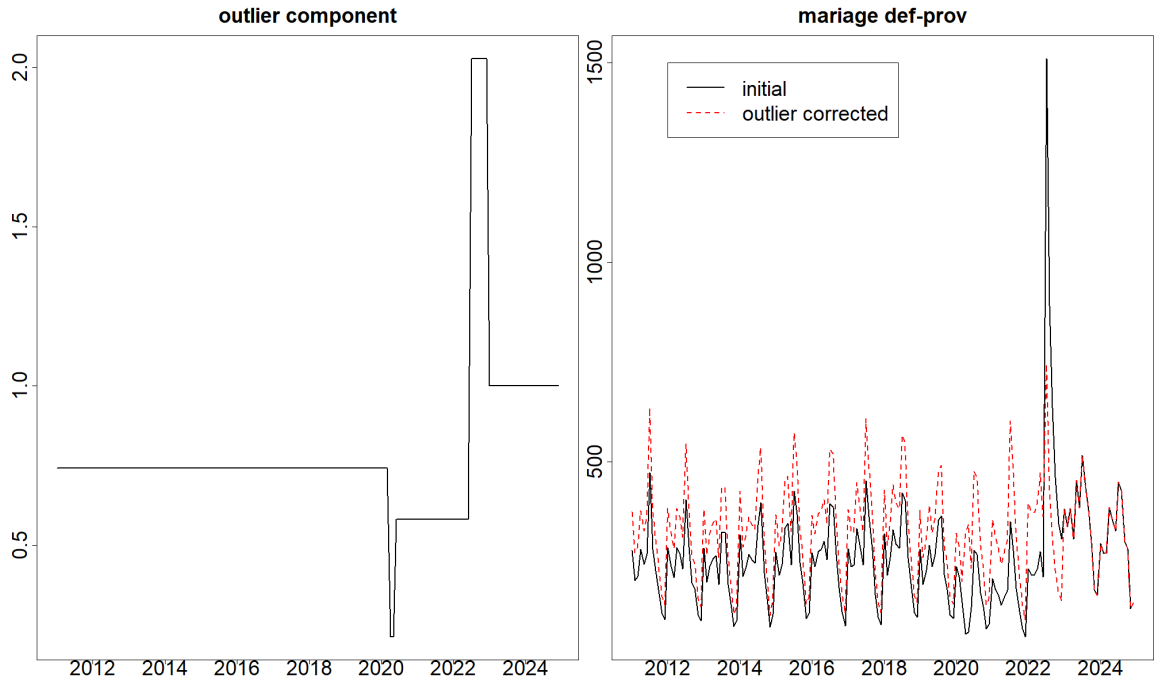


FIGURE 3 À gauche : composante multiplicative liée aux outliers calculée sur la série $\text{diff}_{E,t}$, ou E est ici le mariage. La série corrigée des outliers est égale au ratio entre les données initiales et cette composante multiplicative. À droite : plot de $\text{diff}_{E,t}$ et après correction des outliers.

Sur la base du modèle actuel, un nombre de mariages durablement plus élevé est attendu par rapport aux années 2011 à 2021. Pour être plus précis : la différence entre les données définitives et les données provisoires devrait être globalement plus élevée. Cette hypothèse semble raisonnable à ce stade compte tenu du fait que les données provisoires continuent de ne pas prendre en compte le nombre de mariages de personnes du même sexe, alors que les données définitives, et donc implicitement $\text{diff}_{E,t}$, les prennent en compte. Cependant, cette hypothèse ne tient pas compte de la tendance au ralentissement du nombre de mariages de personnes du même sexe qui se poursuit, même après le Level Shift négatif de janvier 2023. Dans cette optique, les prévisions de nombre total de mariages pour l'année 2025 pourraient être légèrement sur-estimées, voir section 5.

3.3.3 Effets calendrier

Pour les données provisoires, un modèle REG-ARIMA est ajusté avec les outliers trouvés en section 3.3.2, en choisissant l'option 3 dans un premier temps et en laissant X13-ARIMA-SEATS choisir automatiquement le meilleur modèle ARIMA saisonnier. Il en résulte un modèle dont les résidus présentent des effets calendrier, d'après une analyse spectrale. L'option 3 correspond ainsi à un modèle trop simple, qui ne supprime pas correctement tous les effets calendrier. Si la même démarche est appliquée avec l'option 2, les analyses spectrales montrent que les résidus du modèle ne présentent plus d'effet calendrier, de même que la série désaisonnalisée. L'option 2 est ainsi retenue.

Pour la modélisation de $\text{diff}_{E,t}$, les analyses spectrales avec l'option 3 concluent que les résidus du modèle REG-ARIMA ont des effets calendrier tout juste assez faibles pour ne pas être considérés comme significatifs. Cette conclusion est cependant assez peu satisfaisante. S'il est considéré que les jours de la semaine ont une influence différente sur la série des données

provisoires, il est assez logique que cela s'applique aussi sur $\text{diff}_{E,t}$. Ainsi, compte tenu du choix effectué pour les données provisoires et du fait que l'option 3 mène à des résidus dont les effets calendrier sont presque significatifs, l'option 2 est choisie.

3.3.4 Modèle ARIMA et conclusions

X13-ARIMA-SEATS propose un choix de modèle ARIMA saisonnier en se basant sur des critères du type Akaike. Pour les données provisoires, la proposition est un modèle ARIMA(1,1,1)(0,1,1). Les résidus ne montrent pas d'autocorrélations significatives, ce qui permet une validation définitive de ce modèle ARIMA. Le modèle REG-ARIMA final des données provisoires a ainsi 10 outliers, 7 paramètres pour les effets calendrier et 4 paramètres pour la partie ARIMA, ce qui fait tout de même 21 paramètres.

Une réduction du nombre de paramètres pourrait être souhaitable, mais n'est pas facile à réaliser. Les tests ont montré que le nombre de paramètres pour les effets calendrier ne peut pas être facilement réduit en utilisant l'option 3, puisqu'un tel modèle laisse des effets calendrier dans les résidus. Le nombre important d'outliers, 10, pourrait éventuellement être réduit, peut-être en traitant conjointement les dates insolites comme évoqué précédemment. Cependant une telle opération n'est pas aisée, puisque justement les dates insolites n'ont pas toutes le même effet sur le nombre de mariages. En regardant les coefficients β_i associés aux 5 outliers liés à des dates insolites, il est possible de noter que le coefficient de décembre 2012 est bien plus élevé que celui de février 2022 qui est lui même plus élevé que ceux de novembre 2011, février 2020 et mars 2023. À ce stade, il semble raisonnable de valider le modèle ainsi. Les analyses des révisions, qui testent la stabilité du modèle dans le temps en effectuant des modélisations sur une partie des données et non toutes les données, sont concluantes. Les ajustements de modèles sur des sous-ensembles des données donnent des valeurs relativement stables pour les différents paramètres du modèle, voir annexe A.

Pour la série $\text{diff}_{E,t}$, X13-ARIMA-SEATS choisit automatiquement un modèle ARIMA(3,1,1)(0,1,1), après avoir initialement considéré comme candidat un modèle plus simple et bien connu : le modèle "airline", ARIMA(0,1,1)(0,1,1), voir par exemple Brockwell & Davis [2009]. Les p-valeurs associées aux trois coefficients autorégressifs du modèle ϕ_1, ϕ_2, ϕ_3, voir équation (2), sont peu convaincantes et indiquent que certains paramètres ne sont pas significativement différents de zéro. Ceci incite à vérifier si le modèle peut être simplifié. Visuellement, les résidus obtenus avec un modèle airline, présentent des autocorrélations moins fortes que ceux obtenus avec le modèle ARIMA(3,1,1)(0,1,1), ce qui pourrait inciter à choisir le modèle airline.

Le modèle airline n'est cependant pas retenu en raison d'un risque qui est également présent au sein de la série $\text{diff}_{E,t}$ des divorces. Choisir un modèle airline revient à accepter d'avoir des prédictions qui intègrent une tendance basée sur les variations récentes, voir Box & Jenkins [2015], section 9.3.4. Concrètement, le modèle airline tend à extrapoler dans les prédictions l'éventuelle tendance détectée sur les dernières données, et cette tendance extrapolée est linéaire. La fonction de prédiction présentée dans Box & Jenkins [2015], chapitre 9, s'écrit sous la forme $b_0 + b_1.h$, où h est le nombre de pas de temps (mois) de la prédiction. Une telle propriété ne pose pas de soucis pour une série avec une tendance claire, puisque celle-ci est simplement extrapolée. Cependant la série $\text{diff}_{E,t}$ étudiée, corrigée de ses outliers, ne présente pas vraiment de tendance claire, voir figure 2, courbe rouge pointillée sur le graphique de droite. Il n'est donc pas souhaitable d'avoir potentiellement une tendance dans les prédictions.

La solution naturelle pour "amortir" la tendance linéaire de la fonction de prédiction du modèle airline est de rajouter un terme auto-régressif d'ordre 1, $|\phi_1| < 1$, et ainsi de choisir un modèle

ARIMA(1,1,1)(0,1,1) au lieu du modèle airline ARIMA(0,1,1)(0,1,1). Avec ce choix, Box & Jenkins [2015] montrent que la fonction de prédiction s'écrit sous la forme $b_0 + b_1.\phi_1^h$ et ne contient plus de tendance linéaire du fait du facteur d'amortissement $\phi_1^h, |\phi_1| < 1$. Le choix final décrit ici, un modèle ARIMA(1,1,1)(0,1,1), est aussi appliqué pour la série $\text{diff}_{E,t}$ des divorces et vient ainsi éviter de créer des tendances non-souhaitées dans les prédictions. Pour la série $\text{diff}_{E,t}$ des mariages, le modèle final a ainsi 4 outliers, 7 paramètres pour les effets calendrier et 4 paramètres pour la partie ARIMA, ce qui fait 15 paramètres au total.

3.4 Modélisation des divorces

3.4.1 Imputations

La série des divorces est la seule qui commence en janvier 2013. Les trois autres séries étudiées dans ce rapport démarrent en janvier 2011. Pour les divorces, la valeur provisoire de décembre 2013 est manquante et doit être imputée. Par ailleurs, les valeurs provisoires d'octobre et novembre 2017 doivent être imputées, comme expliqué en section 2.3. Ce dernier point n'est pas forcément visible à l'œil nu si les données provisoires sont observées graphiquement. Le souci avec les valeurs provisoire d'octobre et novembre 2017 pour les divorces est cependant plus clair en observant la différence entre les données définitives et provisoires, voir figure 4, graphique de droite. La valeur définitive de novembre 2017 est dans un ordre de grandeur habituel

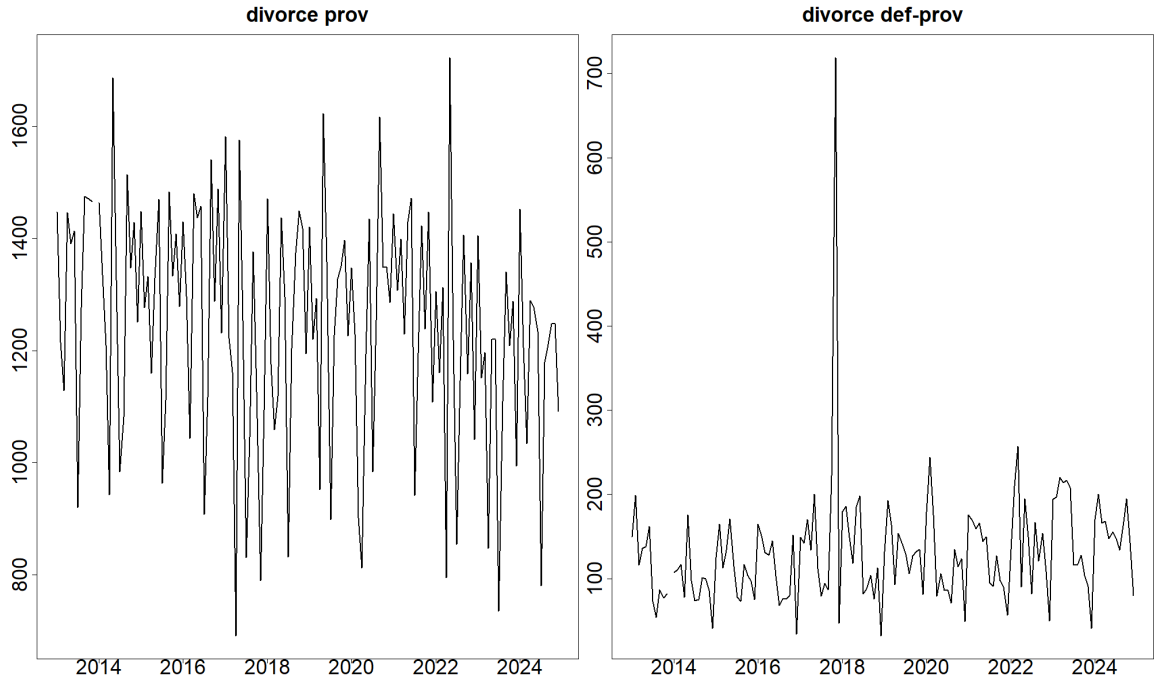


FIGURE 4 Nombre mensuel de divorces en Suisse : données provisoires (à gauche) et différence entre les données définitives et les données provisoires (à droite).

pour un mois de novembre. En revanche, c'est bien la valeur provisoire de ce mois qui est nettement plus faible que d'habitude, avec une valeur de 790 divorces provisoires, alors que pour les mois de novembre de toutes les autres années les données provisoires sont comprises entre 1248 et 1488 divorces. L'effet pour octobre 2017 est plus discret mais tout de même détectable. La valeur provisoire d'octobre 2017 est de 1'127, une valeur assez faible pour un mois d'octobre, sans toutefois être clairement fausse sans information supplémentaire. La valeur définitive permet en revanche de calculer une différence de 222 divorces entre les données définitives et

provisoires, ce qui est plus du double de la moyenne de la série pour un mois d'octobre.

Pour imputer, les observations manquantes ou incomplètes sont remplacées par la valeur 9'999 dans la série $\text{diff}_{E,t}$, ou E désigne désormais les divorces. En utilisant l'option 3 pour traiter les effets calendrier, ce qui sera l'option retenue in fine, X13-ARIMA-SEATS ne détecte pas d'outlier en plus des trois valeurs 9'999 imputées. Le modèle ARIMA proposé est un modèle airline, ce qui est un modèle proche de celui choisi par la suite. Sur la base de ces choix et d'une décomposition qui isole la composante outliers, voir équation (1), la valeur manquante pour $\text{diff}_{E,t}$ est remplacée par la valeur 35. Les valeurs initiales de 222 et 719 pour $\text{diff}_{E,t}$ en octobre et novembre 2017 sont remplacées par des imputations à 100 et 110 respectivement. Cette première imputation est très peu modifiée avec le choix de modèle final : les imputations finales pour $\text{diff}_{E,t}$ en décembre 2013, octobre 2017 et novembre 2017 sont respectivement de 36, 95 et 107. Les données ainsi imputées forment de nouvelles séries $\text{prov}_{E,t}$ et $\text{diff}_{E,t}$ représentées en figure 5.

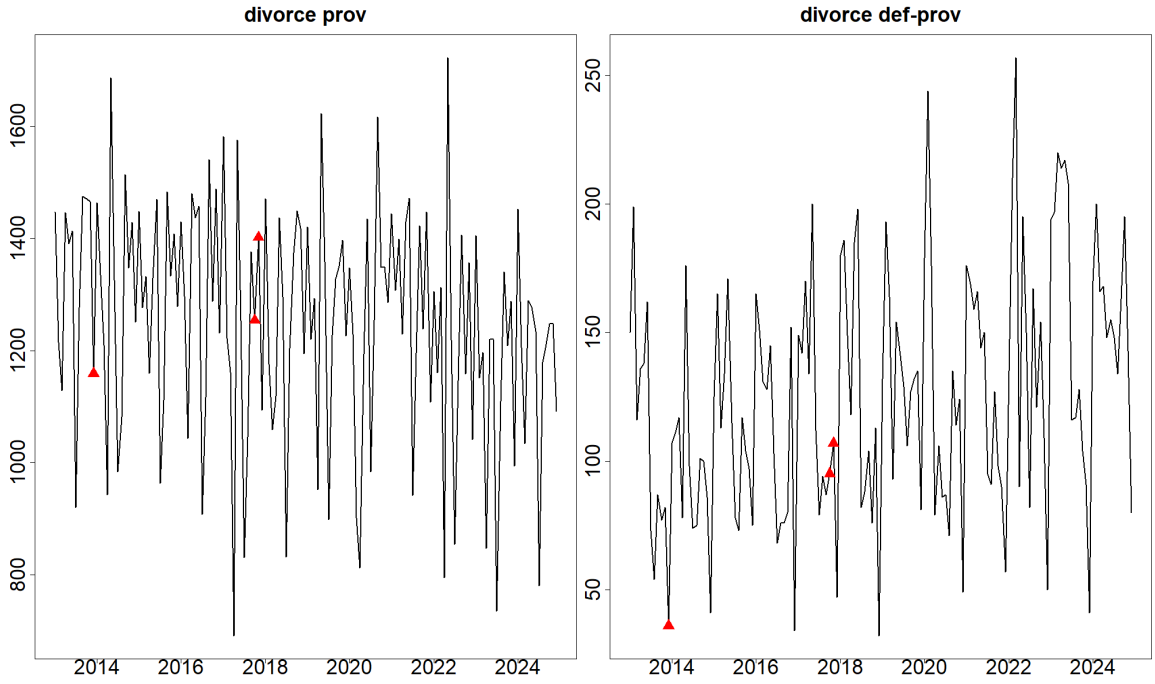


FIGURE 5 Nombre mensuel de divorces en Suisse : données provisoires à gauche et différence entre les données définitives et les données provisoires à droite. Les données imputées à l'aide du modèle final sur $\text{diff}_{E,t}$ sont indiquées par les triangles rouges.

3.4.2 Outliers

La détection automatique d'outliers est effectuée sur les données provisoires en utilisant l'option 2 pour les effets calendrier, qui est l'option adoptée in fine. Si le seuil de détection d'outlier est laissé à sa valeur par défaut, X13-ARIMA-SEATS choisit un modèle ARIMA airline, et détecte deux AOs en avril 2017 et avril 2022 avec des coefficients β_i négatifs, ce qui correspond à un nombre de divorces particulièrement bas. Ces deux observations ne peuvent cependant pas être interprétées comme des outliers dans le sens où aucun évènement exceptionnel ou inattendu ne vient perturber à la hausse ou à la baisse ces deux observations.

L'interprétation la plus probable pour les valeurs faibles en avril 2017 et avril 2022 est l'existence de jours fériés judiciaires. Concrètement, les tribunaux qui prononcent les divorces sont

généralement fermés toute la semaine du vendredi saint et toute la semaine du lundi de Pâques, soit deux semaines consécutives de fermeture. En 2017 et 2022, le dimanche de Pâques tombe respectivement le 16 et 17 avril. Les jours fériés judiciaires des années 2017 et 2022 sont ainsi intégralement concentrés sur le mois d'avril. Les faibles valeurs observées en avril 2017 et avril 2022 relèvent donc probablement d'un effet calendrier et en particulier un effet Pâques. Un tel effet peut être pris en compte dans X13-ARIMA-SEATS à l'aide de prédicteurs spécifiques dans la partie REG du modèle REG-ARIMA, puisque le logiciel propose notamment un prédicteur nommé "easter".

Dans le présent travail, le choix actuel est de ne pas utiliser ces prédicteurs prédéfinis et de se fier seulement à un calendrier Suisse des jours fériés. Ainsi, pour le moment, cet effet Pâques n'est pas parfaitement pris en compte. Ceci n'est pas considéré comme un problème majeur dans la mesure où le modèle pour les données provisoires ne cherche jamais à prédire une valeur inconnue au mois d'avril. L'utilisation actuelle de ce modèle est seulement focalisée sur une prédiction au mois de décembre. Une prise en compte plus fine de l'effet Pâques a toutefois le potentiel d'améliorer globalement le modèle et constitue une perspective d'amélioration future.

Une solution pour ne pas accepter avril 2017 et avril 2022 comme outlier pour les données provisoires de divorce est de relever le seuil de détection "outlier.critical". Si celui-ci est relevé à 5 au lieu de sa valeur par défaut, ces dates cessent d'être considérées comme des outliers. Le modèle retenu pour les données provisoires des divorces ne contient ainsi aucun outlier. Pour la différence $\text{diff}_{E,t}$, X13-ARIMA-SEATS ne détecte aucun outlier, même en utilisant le seuil par défaut. Ainsi, aucun des deux modèles utilisés pour modéliser les divorces ne contient d'outlier.

3.4.3 Effets calendrier

Pour les données provisoires, un ajustement de modèle avec X13-ARIMA-SEATS sans outliers est testée en utilisant l'option 3 pour les effets calendrier, c'est-à-dire un modèle relativement simple à 3 paramètres, voir section 3.2.2. L'ajustement donne un modèle ARIMA airline, ce qui sera le modèle définitif. Cependant, avec l'option 3, l'analyse du spectre de la série désaisonnalisée montre des effets calendrier résiduels très importants, qui sont communément appelés effets "trading day". Ceci montre que l'option 3 correspond à un modèle trop simple pour supprimer les effets calendrier, et que le modèle doit être affiné, par exemple en laissant les jours non-fériés avoir une influence différente. L'utilisation de l'option 2, mène aussi à un choix de modèle airline pour la partie ARIMA. Les analyses spectrales ne montrent plus d'effet trading day sur les résidus du modèle ou sur la série désaisonnalisée. L'option 2 est ainsi retenue. Comme discuté en section précédente, une amélioration possible serait de capter l'effet lié à la date du dimanche de Pâques sur les données.

Pour la série $\text{diff}_{E,t}$, un ajustement sans outliers avec l'option 3 mène aussi à un modèle airline pour la partie ARIMA. Les analyses spectrales ne montrent cette fois-ci pas d'effets calendrier résiduels. L'option 3 est ainsi retenue, même s'il s'agit d'une option différente de celle choisie pour la série $\text{prov}_{E,t}$.

3.4.4 Modèle ARIMA et conclusions

Pour chacune des deux séries, un modèle airline, ARIMA(0,1,1)(0,1,1) est choisi par X13-ARIMA-SEATS. Il s'agit donc de savoir si ce modèle peut être validé. Pour les données provisoires, les résidus du modèle ne présentent pas d'autocorrélations significatives. Les analyses de révision de tous les paramètres du modèle montrent une bonne stabilité de celui-ci dans le temps, ce qui

permet d'accepter le modèle airline. Le risque de tendance dans les prédictions présenté en section 3.3.4 ne s'applique pas à la série des données provisoires des divorces pour deux raisons. D'une part, la série présente une légère tendance à la baisse et il n'est pas gênant d'extrapoler cette tendance dans les prédictions. D'autre part, la prédiction effectuée avec ce modèle porte seulement sur un horizon de temps d'un mois puisque seule la valeur provisoire de décembre est prédite. Le modèle airline, ARIMA(0,1,1)(0,1,1), est ainsi validé. Finalement, le modèle des divorces provisoires n'a pas d'outlier, comporte 7 paramètres pour les effets calendrier et 3 paramètres pour la partie ARIMA, soit 10 paramètres au total.

Pour la série $\text{diff}_{E,t}$, le modèle airline est proposé par X13-ARIMA-SEATS et les résidus présentent des autocorrélations assez fortes aux lags 4, 9 et 10. Pour les raisons détaillées en section 3.3.4, le modèle airline ne peut pas être conservé. Le choix d'un modèle ARIMA(1,1,1)(0,1,1) ne résout pas le problème des autocorrélations fortes mais donne un coefficient autorégressif ϕ_1 significativement différent de zéro. Tout comme pour la série $\text{diff}_{E,t}$ des mariages, le modèle ARIMA(1,1,1)(0,1,1) est préféré au modèle airline ARIMA(0,1,1)(0,1,1). Le modèle final n'a pas d'outliers, comporte 3 paramètres pour les effets calendrier et 4 pour la partie ARIMA soit 7 paramètres.

3.5 Modélisation des naissances

3.5.1 Imputations

La série des naissances commence en janvier 2011 et, tout comme pour les mariages, les données provisoires de décembre 2011, 2012 et 2013 sont manquantes et doivent être imputées. La stratégie d'imputer la série $\text{diff}_{E,t}$, où E désigne maintenant les naissances, s'applique. L'option utilisée pour les effets calendrier est l'option 3, c'est-à-dire celle qui sera finalement choisie. Avec cela, X13-ARIMA-SEATS choisit un modèle ARIMA(0,1,2)(0,1,1), ce qui ne sera pas le modèle définitif. Un outlier en avril 2014 est détecté pour $\text{diff}_{E,t}$ et celui-ci ne sera pas retenu par la suite, ni pour le modèle sur les données provisoires, ni pour le modèle sur $\text{diff}_{E,t}$. Sur la base de ce modèle, les imputations pour $\text{diff}_{E,t}$ en décembre 2011, 2012 et 2013 sont respectivement de 220, 247 et 246. Si le modèle final est utilisé pour produire les imputations, ces 3 valeurs changent peu et sont respectivement de 217, 243 et 229. Ce sont ces dernières valeurs qui sont conservées. Les données livrées par la section DEM sont représentées en figure 6 et les données complétées par les imputations en figure 7.

3.5.2 Outliers

La détection d'outlier pour les données provisoires donne des signaux légèrement contradictoires, avec parfois aucun outlier et parfois des outliers qui dépassent tout juste le seuil de détection, selon le modèle utilisé. Contrairement à l'intuition, les données provisoires sur les naissances montrent de forts effets calendrier pour des raisons d'enregistrement. Dans les données, l'enregistrement des naissances se fait avec un retard qui peut dépendre du calendrier. Ainsi, pour la détection automatique d'outliers par X13-ARIMA-SEATS dans les données provisoires, c'est l'option 2 qui est utilisée pour modéliser les effets calendrier. Une détection automatique avec cette option et en laissant le choix du modèle ARIMA ne conduit, avec le seuil par défaut, à aucun outlier. Le modèle ARIMA choisi est un ARIMA(3,1,1)(0,1,1) et il apparaît que ce choix est peu convaincant, avec un coefficient ϕ_2 quasi nul et un coefficient ϕ_3 un peu plus significatif, mais seulement égal à 2 écart-types ce qui reste faible. Ce modèle ARIMA utilisé pour conclure qu'il n'y a pas d'outliers n'est pas le modèle finalement choisi. Si c'est le modèle ARIMA final qui est utilisé pour détecter automatiquement des outliers, X13-ARIMA-SEATS détecte un Level

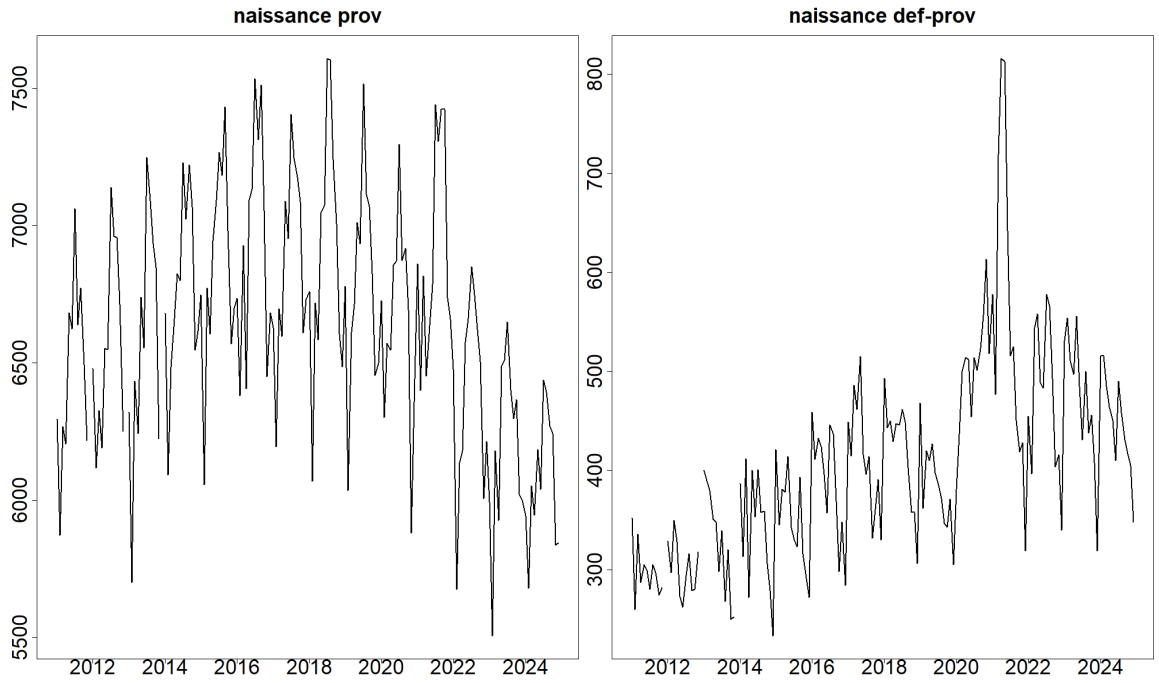


FIGURE 6 Nombre mensuel de naissances en Suisse : données provisoires (à gauche) et différence entre les données définitives et les données provisoires (à droite).

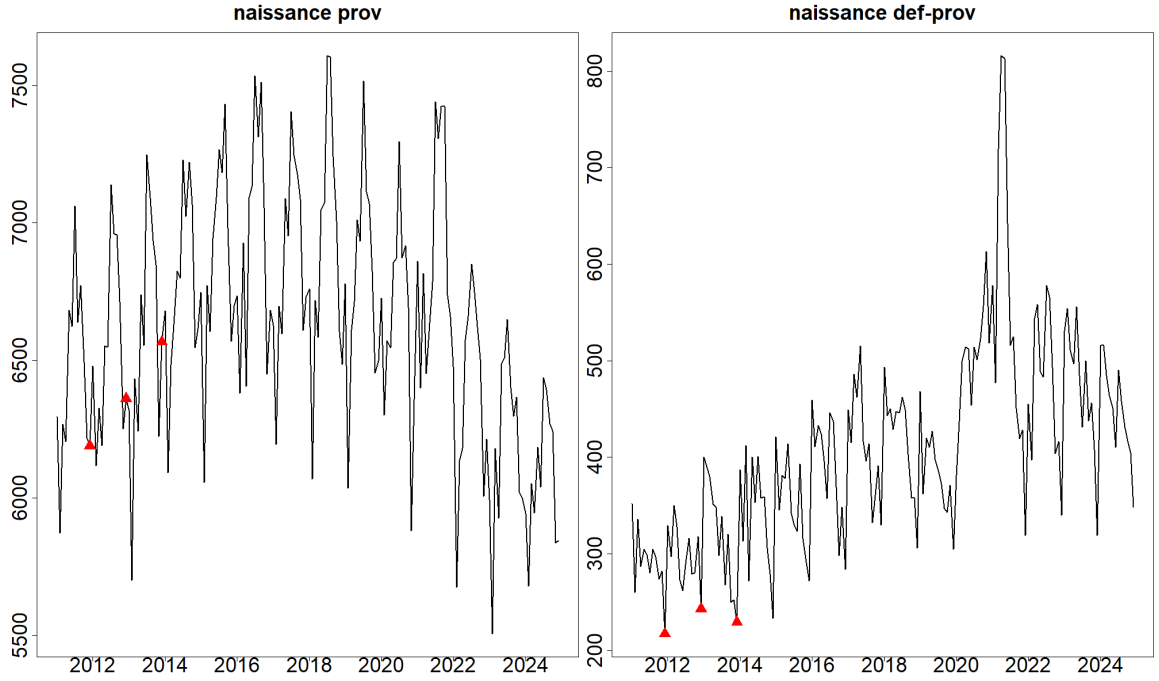


FIGURE 7 Nombre mensuel de naissances en Suisse : données provisoires à gauche et différence entre les données définitives et les données provisoires à droite. Les données imputées à l'aide du modèle final sur $\text{diff}_{E,t}$ sont indiquées par les triangles rouges.

Shift (LS) positif en janvier 2021. L'amplitude de ce LS est tout juste suffisante pour que l'observation soit sélectionnée par X13-ARIMA-SEATS. La date de janvier 2021 tombe effectivement à peu près 9 mois après le début du premier confinement lié à la pandémie, connu en Suisse pour avoir causé un léger baby boom. Toutefois, l'effet est faible et un relèvement marginal du seuil de détection à 4 fait disparaître cet outlier. Pour ces raisons, il est décidé pour le moment de travailler sans outlier sur les données provisoires et donc de ne pas retenir ce LS en janvier 2021.

Pour la série $\text{diff}_{E,t}$, comme évoqué à la section précédente, un AO est détecté en avril 2014, avec un coefficient β_i négatif qui reflète une observation en avril 2014 plus faible qu'attendu. Cet outlier est détecté en laissant X13-ARIMA-SEATS choisir le modèle ARIMA. Le modèle ARIMA choisi n'est pas le modèle finalement retenu, voir section 3.5.4. Si le modèle ARIMA final est imposé pour détecter automatiquement les outliers, l'observation d'avril 2014 cesse d'être un outlier. X13-ARIMA-SEATS mentionne tout de même que cette observation est "presque" un outlier, dans le sens où elle se situe juste en dessous du seuil de détection par défaut. Compte tenu de l'absence d'interprétation claire et du fait que l'observation d'avril 2014 est, selon le modèle ARIMA utilisé, juste au-dessus ou en dessous du seuil de détection, il n'y a pas lieu de retenir avril 2014 comme un outlier. Ainsi la série $\text{diff}_{E,t}$, tout comme la série des données provisoires, est considérée comme sans outliers.

3.5.3 Effets calendrier

Avec X13-ARIMA-SEATS, l'option 3 associée à un ajustement de modèle sans outliers pour les données provisoires mène à un modèle ARIMA(3,1,1)(0,1,1) assez peu convaincant, comme discuté en section 3.5.2. Un tel modèle montre immédiatement des effets "trading day" dans les résidus, ce qui indique que l'option 3 ne convient pas telle quelle pour traiter les effets calendrier. Si le modèle ARIMA(3,1,1)(0,1,1) est remplacé par un modèle ARIMA(1,1,1)(0,1,1), c'est-à-dire par le modèle final, l'effet trading day sur les résidus persiste, ce qui permet d'effacer complètement l'option 3. L'utilisation de l'option 2 sans outliers et en laissant X13-ARIMA-SEATS choisir le modèle ARIMA mène à un modèle ARIMA(1,1,1)(0,1,1) dont les résidus n'ont pas d'effet trading day. La série désaisonnalisée présente cependant encore des effets trading day d'après l'analyse de son spectre. Ce problème ne se résout pas si l'option 1, qui donne une influence différente aux différents jours fériés selon qu'ils tombent un lundi, mardi, etc., est utilisée. Au contraire, les analyses spectrales avec l'option 1 s'avèrent moins bonnes qu'avec l'option 2, ce qui peut sembler contre-intuitif mais montre aussi les limites de ces analyses spectrales. Il apparaît au final que l'option 2 est le meilleur compromis pour modéliser les effets calendrier sur les données provisoires des naissances.

Pour la différence $\text{diff}_{E,t}$, l'option 3 donne des résultats satisfaisants. Le choix automatique de modèle ARIMA mène à un modèle satisfaisant au niveau de l'analyse spectrale des résidus, mais qui ne peut pas être conservé en l'état, voir section 3.5.4. Si c'est le modèle ARIMA définitif qui est utilisé avec l'option 3, l'analyse spectrale se dégrade légèrement, avec une série désaisonnalisée qui montre un pic dans son spectre à l'une des fréquences "trading day". Cependant, ce léger problème ne se résout pas en choisissant l'option 2 à la place de l'option 3. L'option 3 est ainsi le meilleur choix pour modéliser les effets calendrier de la différence $\text{diff}_{E,t}$.

3.5.4 Modèle ARIMA et conclusions

Le modèle définitif pour les données provisoires des naissances est celui suggéré par X13-ARIMA-SEATS lorsqu'aucun outlier n'est choisi et que l'option 2 est utilisée pour les effets calendrier. Il s'agit d'un modèle ARIMA(1,1,1)(0,1,1) soit un modèle un peu plus complexe que le

modèle airline puisqu'il comporte en plus un paramètre autorégressif ϕ_1. Les résidus du modèle ne présentent pas d'autocorrélations significatives, ce qui est satisfaisant. Le modèle final a ainsi 7 paramètres pour les effets calendrier, 4 paramètres pour la partie ARIMA et aucun outlier, soit 11 paramètres au total.

Pour la différence $\text{diff}_{E,t}$, un ajustement automatique via X13-ARIMA-SEATS sans outliers et avec l'option 3 suggère aussi un modèle ARIMA(1,1,1)(0,1,1). Cependant, un tel modèle mène à des résidus ayant une autocorrélation significativement grande (3.5 écart-types) au lag 9, c'est-à-dire que les résidus du modèle séparés de 9 mois sont corrélés de manière significative. En l'occurrence, la corrélation constatée est une corrélation négative, qui dépasse largement les seuils de significativité classiques situés à 1.96 ou 2 écart-types. Un modèle ARIMA qui laisse des autocorrélations significatives dans les résidus ne devrait normalement pas être accepté, l'objectif des modèles ARIMA étant précisément de capter la structure d'autocorrélation présente dans les séries temporelles. L'originalité ici est que la forte autocorrélation dans les résidus se situe au lag 9 mois, sans qu'il y ait d'autocorrélation résiduelle significative aux lag inférieurs, 1 mois, 2 mois, etc. L'ajout d'un terme autorégressif supplémentaire via un modèle ARIMA(2,1,1)(0,1,1) mène à un modèle qui présente toujours cette autocorrélation résiduelle. L'ajout d'un terme "moving average" via un modèle ARIMA(1,1,2)(0,1,1) ne fonctionne pas mieux. Dans chaque cas, le paramètre ajouté est peu significatif car proche de son écart-type et le problème persiste. Un travail supplémentaire sur la partie saisonnière du modèle ne peut par définition que traiter les problèmes d'autocorrélations sur des lags multiples de 12, ce qui n'est pas le cas ici.

La méthode retenue pour traiter le problème d'autocorrélation est d'ajouter au modèle un terme "moving average" spécifique au lag 9, noté classiquement θ_9, sans utiliser de termes aux lags intermédiaires 2, 3, ..., 8. La partie moving average du modèle ARIMA comporte ainsi un paramètre θ_1 au lag 1, un paramètre θ_9 au lag 9, mais pas de paramètres $\theta_2, \theta_3, \ldots, \theta_8$. De tels modèles sont utilisés avec X13-ARIMA-SEATS via l'instruction :
`arima.model=''(1 1 [1 9])(0 1 1)''`.
Dans le modèle estimé, le paramètre θ_9 vaut plus de 3 fois son écart-type et est donc largement significatif. Les résidus d'un tel modèle ne présentent plus d'autocorrélations significatives. Ce problème d'autocorrélation négative des résidus au lag 9 mois est relativement difficile à interpréter. La principale difficulté vient du fait que ceci est constaté pour la série $\text{diff}_{E,t}$, mais pas pour la série des données provisoires. Le fait qu'un mois "fort" en termes de naissances tend à créer un mois "faible" 9 mois plus tard est interprétable, puisqu'une femme peut difficilement avoir des enfants de deux grossesses séparées à 9 mois d'intervalle ou moins. Cependant, ceci n'explique pas pourquoi le phénomène est constaté seulement sur $\text{diff}_{E,t}$. Pour le moment, il est décidé d'intégrer ce paramètre θ_9 au modèle, en laissant toujours la possibilité d'amélioration du modèle à l'avenir. Le modèle retenu finalement n'a pas d'outlier, comporte 3 paramètres pour les effets calendrier et 5 paramètres pour le modèle ARIMA, soit 8 paramètres au total.

3.6 Modélisation des décès

3.6.1 Imputations

Les données provisoires des décès vont de janvier 2011 à décembre 2024 et présentent trois valeurs manquantes et deux valeurs incomplètes. Les valeurs manquantes sont les mêmes que pour les autres séries étudiées ici. Les valeurs incomplètes sont celles de novembre et décembre 2024, soit les deux dernières valeurs de la série provisoire. Le souci est lié à des problèmes d'enregistrements et n'est pas facilement visible en regardant simplement graphiquement les données provisoires. Cependant, la figure 8, graphique de droite, donne la courbe de

diff$_{E,t}$ où E désigne maintenant les décès et montre plus clairement un problème en novembre et décembre 2024. Toutes les imputations sont effectuées selon les méthodes décrites précé-

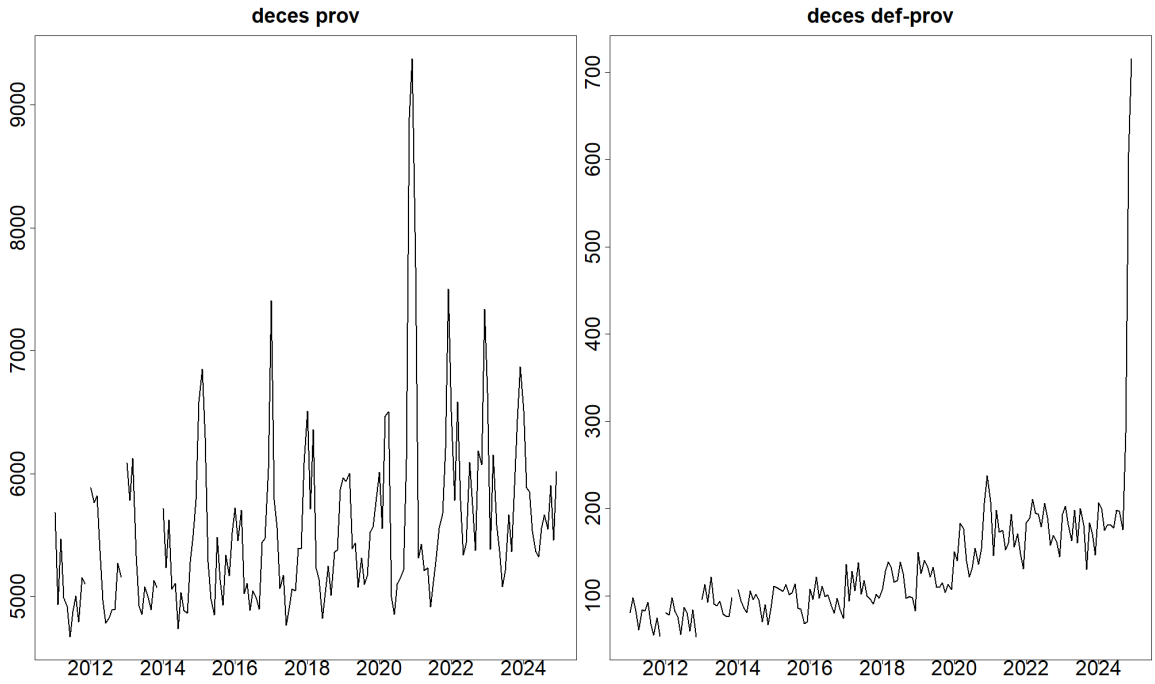


FIGURE 8 Nombre mensuel de décès en Suisse : données provisoires (à gauche) et différence entre les données définitives et les données provisoires (à droite).

demment, c'est-à-dire à l'aide d'un modèle sur diff$_{E,t}$. Une évaluation avec X13-ARIMA-SEATS avec l'option 3 et en laissant celui-ci chercher les outliers et décider du modèle ARIMA mène à un modèle airline et des outliers en novembre 2020, décembre 2020 et octobre 2024. Sur la base de la décomposition de l'équation (1), les imputations pour les trois valeurs manquantes sont 56, 65 et 73. Les imputations pour les deux dernières observations sont de 163 et 191. Si le modèle final de la section 3.6.4 est utilisé pour recalculer les imputations, les changements sont très limités. Les trois valeurs manquantes sont remplacées par 57, 65 et 72. Les deux dernières observations sont remplacées par 181 et 169. Les séries enrichies des valeurs imputées sont représentées en figure 9.

3.6.2 Outliers

Le choix des outliers pour les données provisoires dépend du seuil de détection. En laissant celui-ci à sa valeur par défaut, 7 outliers sont détectés par X13-ARIMA-SEATS si l'option 3 est utilisée pour les effets calendrier et si X13-ARIMA-SEATS choisi lui même le modèle ARIMA. Ces 7 outliers peuvent être séparés en deux groupes. Un total de 5 AO sont qualifiés de forts car le coefficient β_i associé est très élevé en valeur absolue. Deux outliers, plus faibles, sont reliés à des coefficients β_i dont la t-valeur est tout juste supérieure à 4, c'est-à-dire proche du seuil de détection par défaut. Les 5 outliers forts se produisent en janvier 2017, avril 2020, novembre 2020, décembre 2020 et janvier 2021. Leur interprétation est claire, et ces outliers sont ajoutés au modèle.

- La forte mortalité de janvier 2017 est liée à une importante épidémie de grippe saisonnière qui a duré de mi-décembre 2016 à mi-février 2017, avec justement un pic en janvier.
- La forte mortalité d'avril 2020 est due à la première vague de contamination dans la pandémie de COVID-19.

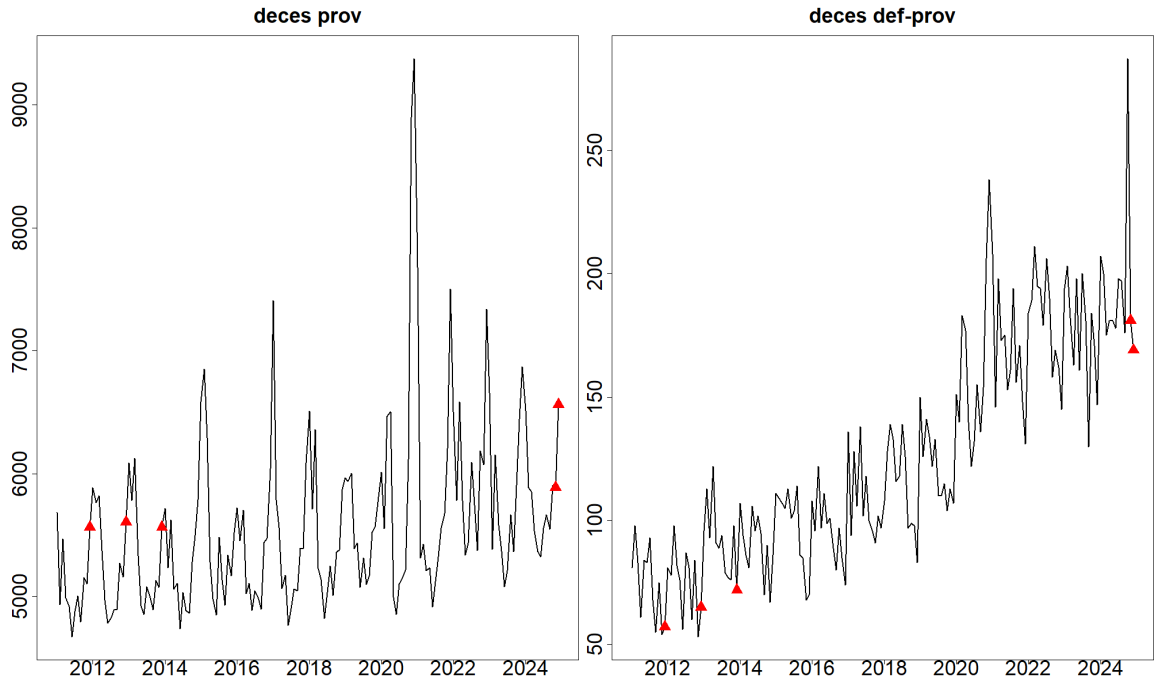


FIGURE 9 Nombre mensuel de décès en Suisse : données provisoires à gauche et différence entre les données définitives et les données provisoires à droite. Les données imputées à l'aide du modèle final sur $\text{diff}_{E,t}$ sont indiquées par les triangles rouges.

— La forte mortalité de novembre 2020 à janvier 2021 est due à la seconde vague. Les observations de novembre 2020 et décembre 2020 sont particulièrement fortes, avec des t-valeurs supérieures à 10. L'observation de janvier 2021 reste un outlier fort avec une t-valeur proche de 6.

Les deux outliers plus faibles détectés si le seuil par défaut est utilisé se produisent en février 2015 et décembre 2021, avec toujours une forte mortalité et non l'inverse. Les t-valeurs sont proches de 4. Ces outliers sont interprétables, avec des interprétations similaires à celles des outliers forts : grippe pour l'année 2015 et COVID-19 pour la fin de l'année 2021, mais avec une amplitude plus faible que pour la fin 2020. Pour le moment il est décidé de ne pas conserver les deux outliers les plus faibles, ce qui revient implicitement à utiliser un seuil de détection supérieur à 4, par exemple 4.5. Ce choix pourrait cependant être reconsidéré à l'avenir. Les outliers finalement retenus sont ainsi au nombre de 5 et sont tous des additive outliers.

Pour la série $\text{diff}_{E,t}$, où E désigne les décès, la détection automatique d'outliers conduit à un seul AO en décembre 2020, clairement interprétable compte tenu des explications précédentes. Ceci est obtenu avec le seuil par défaut, en utilisant l'option 3 pour les effets calendrier et en laissant X13-ARIMA-SEATS choisir le modèle ARIMA.

3.6.3 Effets calendrier

Le comportement de la population étant en moyenne différent selon qu'il s'agisse d'un jour ouvrable ou non, la série des décès pourrait avoir des effets calendrier importants. Par exemple, selon les statistiques de l'Office Fédéral des Routes [4], la mortalité sur la route est parfois plus élevée le vendredi et le week-end que durant les autres jours de la semaine. Le nombre de morts

4. Statistique des accidents de la route 2025, téléchargeable sur https://www.astra.admin.ch/

sur les routes suisses est cependant plus bas que dans le passé, avec 214 victimes en 2025. Il n'est donc pas évident que les données de mortalité toutes causes confondues montrent des effets calendrier décelables statistiquement. Pour les données provisoires, un ajustement de modèle est effectué avec X13-ARIMA-SEATS en utilisant les 5 outliers explicités en section 3.6.2 et l'option 3 pour les effets calendrier. Le modèle obtenu a de bonnes performances au niveau des analyses spectrales avec une absence d'effet calendrier sur les résidus du modèle ou sur la série désaisonnalisée. Ceci permet de valider l'option 3 qui pour rappel donne la même influence aux jours non-fériés du lundi au vendredi.

La série $\text{diff}_{E,t}$ étant une différence, elle a généralement moins ou autant, mais pas plus, d'effets calendrier que la série de données provisoires, du fait d'un possible effet de compensation des effets calendrier présents dans la série définitive et la série provisoire. Les analyses montrent ainsi des performances satisfaisantes au niveau des analyses spectrales si l'option 3 est utilisée pour le choix de modèle sur $\text{diff}_{E,t}$.

3.6.4 Modèle ARIMA et conclusions

Pour les données provisoires, le choix automatique de X13-ARIMA-SEATS avec les outliers décrits précédemment et l'option 3 est un modèle ARIMA(0,1,2)(0,1,1). Un tel modèle donne des résidus relativement satisfaisants en termes d'autocorrélation, avec tout de même une autocorrélation négative, de plus de 2 écart-types en valeur absolue, relevée au lag 12. L'autocorrélation des résidus est également assez forte (juste en dessous de 2 écart-types) au lag 24. Ceci pousse à rechercher d'éventuelles améliorations sur la partie saisonnière du modèle ARIMA, par exemple en testant celui-ci avec un paramètre supplémentaire et donc en testant les performances d'un ARIMA(0,1,2)(0,1,2) et d'un ARIMA(0,1,2)(1,1,1). Chacune de ces deux possibilité donne un modèle dont le paramètre supplémentaire est significativement différent de zéro, et donc jugé comme utile. De plus, dans chaque cas, les autocorrélations résiduelles fortes constatées aux lag 12 et 24 disparaissent. Le choix final se porte sur le modèle ARIMA(0,1,2)(1,1,1) qui donne des résultats légèrement meilleurs en termes d'autocorrélation des résidus. Le modèle final des données provisoires des décès a ainsi 5 outliers et utilise l'option 3 pour modéliser les effets calendrier. Le modèle ARIMA est un ARIMA(0,1,2)(1,1,1), ce qui fait au total 5 + 3 + 5 = 13 paramètres.

Pour la série $\text{diff}_{E,t}$, le choix de X13-ARIMA-SEATS se porte sur un modèle airline. Les résidus ont des autocorrélations relativement élevées aux lag 2 et 3, mais qui restent sous le seuil de significativité. Les tentatives d'améliorer cette situation en ajoutant un paramètre à la partie non-saisonnière du modèle ne fonctionnent pas et mènent chaque fois à un paramètre supplémentaire dont la valeur n'est pas significativement différente de zéro. Le modèle airline est ainsi validé pour la partie ARIMA. Avec un seul outlier et l'option 3 pour les effets calendrier, le modèle final comporte 7 paramètres.

3.7 Synthèse sur les modèles et améliorations possibles

Un récapitulatif des choix de modèles est donné dans les tables 1 et 2.

Voici un résumé synthétique des possibles changements, évolutions, améliorations qui pourraient être apportés aux modèles.

- Mariages : les mariages restent l'évènement le plus difficile à prédire avec précision, comme présenté en section 4. Les modèles pour $\text{prov}_{E,t}$ et $\text{diff}_{E,t}$, où E désigne le mariage, sont pourtant les deux modèles avec le plus de paramètres parmi les 8 modèles

Table 1 Récapitulatif des choix de modèles REG-ARIMA pour les données provisoires sur les mariages, divorces, naissances et décès.

Données provisoires	Outliers	Effets calendr. option	Modèle ARIMA saisonnier	Nombre de param.
mariages	AO dates insolites : Nov. 2011, Déc. 2012, Jan. 2013, Fév.2020, Fév. 2022, Mar. 2023. AO liés au COVID-19 : Avr. 2020, Mai. 2020, Juin 2020, Oct. 2020.	2	ARIMA (1,1,1)(0,1,1)	21
divorces	aucun	2	ARIMA (0,1,1)(0,1,1)	10
naissances	aucun	2	ARIMA (1,1,1)(0,1,1)	11
décès	AO lié à la grippe saisonnière : Jan. 2017. AO liés au COVID-19 : Avr. 2020, Nov. 2020, Déc. 2020, Jan. 2021.	3	ARIMA (0,1,2)(1,1,1)	13

Table 2 Récapitulatif des choix de modèles REG-ARIMA pour la différence $\text{diff}_{E,t}$ sur les mariages, divorces, naissances et décès.

Données def.-prov.	Outliers	Effets calendr. option	Modèle ARIMA saisonnier	Nombre de param.
mariages	LS liés au COVID-19 : Avr. 2020, Juin 2020. LS liés à l'autorisation de mariage de couples du même sexe : Juillet 2022, Jan. 2023.	2	ARIMA (1,1,1)(0,1,1)	15
divorces	aucun	3	ARIMA (1,1,1)(0,1,1)	7
naissances	aucun	3	ARIMA (1,1,[1,9])(0,1,1)	8
décès	AO lié au COVID-19 : Déc. 2020.	3	ARIMA (0,1,1)(0,1,1)	7

présentés dans ce travail. Il est possible qu'une meilleure gestion des outliers permette d'améliorer le modèle en prenant mieux en compte les dates insolites. L'autre facteur important qui vient diminuer les performances du modèle est lié à la rupture de série de $\text{diff}_{E,t}$ à partir de juillet 2022 en raison de l'autorisation du mariage des personnes du même sexe. Pour le moment, une modélisation ad-hoc de ce phénomène est faite via des Level Shifts sur $\text{diff}_{E,t}$, mais il existe peut-être un potentiel d'amélioration sur ce point.

— Divorces : le principal point d'amélioration potentiel porte sur une modélisation plus fine des effets calendrier de cette série. L'effet Pâques important n'est pas encore correcte-

ment modélisé. Il serait aussi intéressant d'étudier, en plus de l'effet Pâques, un éventuel effet calendrier spécifique au mois de décembre avec d'éventuelles fermetures des tribunaux pour les fêtes de fin d'année. Pour ce travail, ce point est peut-être encore plus important que l'effet Pâques puisque le modèle des données provisoires vise justement à effectuer des prédictions pour la valeur manquante du mois de décembre.

- Naissances : la série des naissances est celle pour laquelle les performances en prédiction sont les meilleures parmi les 4 séries étudiées. Des améliorations sont cependant peut-être toujours possibles, par exemple par une meilleure compréhension du modèle de la série $\text{diff}_{E,t}$ qui contient pour le moment une caractéristique relativement originale avec ce paramètre moving average θ_9, voir section 3.5.4.
- Décès : en dépit des nombreux outliers observés sur la série des décès, les performances en prédiction sont relativement bonnes. Des améliorations pourraient être effectuées par exemple en recherchant une modélisation performante spécifiquement pour la prédiction du mois de décembre. Le mois de décembre est en effet un mois comptabilisant un nombre de décès relativement élevé, sur lequel des outliers sont plus susceptibles de se produire.

4 Performance des modèles en simulation

4.1 Description des simulations effectuées

Afin d'évaluer la performance des méthodes et modèles proposés, des simulations rétrospectives, backtests, sont mis en œuvre. L'objectif de cette approche est d'évaluer la précision des flash estimates sur les années passées, pour lesquelles les données définitives sont connues. Dans le backtest, les flash estimates sont produites en simulant une situation la plus réaliste possible en pratique, c'est-à-dire sans connaître les outliers futurs ou, bien sûr, les données à prédire. Plusieurs années sont sélectionnées comme années cibles. Il s'agit des années 2019 à 2024, soit six années cibles différentes.

Si a est une année cible donnée, pour laquelle des flash estimates sont calculées, seules les informations qui seraient effectivement disponibles au début de l'année $a+1$ sont utilisées, c'est-à-dire les données provisoires jusqu'au mois de novembre de l'année a et les données définitives jusqu'au mois de décembre de l'année $a-1$. Les données définitives de l'année a sont supposées inconnues et ne servent pas à la production des estimations. En revanche, elles sont utilisées pour les calculs de performance, en comparant justement les estimations produites aux données réelles. Les modèles utilisés pour produire les estimations sont ceux résumés en section 3.7. Les ordres p, P, q et Q des modèles ARIMA saisonniers des 8 modèles étudiés sont ainsi connus et les paramètres correspondants sont estimés sur la base des informations connues. Les outliers sont connus lorsqu'ils sont inclus dans les données supposées connues. Par exemple, pour produire les flash estimates des décès de l'année 2020, la valeur provisoire de novembre 2020 est un outlier et cette information est supposée connue. En revanche, le fait que décembre 2020 soit aussi un outlier n'est pas une information connue pour produire ces flash estimates. Ce dernier point est d'ailleurs responsable des performances mitigées observées sur les décès pendant le COVID-19.

La qualité des estimations est évaluée en comparant les estimations obtenues agrégées sur l'année a aux données définitives agrégées sur cette même année. En d'autres termes, le nombre de MDND estimé sur toute l'année a à l'aide des modèles, noté ici $\widehat{E}_a$ où E désigne les mariages, divorces, naissances ou décès, est comparé au nombre de MDND réel sur cette année E_a. Pour

faciliter les comparaisons, une erreur relative exprimée en pour cent se calcule via la formule :

$$\text{RelError}_{E,a} := \left| \frac{\widehat{E}_a}{E_a} - 1 \right| \tag{3}$$

Le backtest est effectué pour plusieurs années cibles a, ce qui permet d'obtenir un ordre de grandeur des erreurs de prévision et d'évaluer la robustesse de la méthode dans différents contextes temporels. L'erreur de prévision, c'est-à-dire la différence entre $\widehat{E}_a$ et E_a, provient de deux sources. D'une part, la différence $\text{diff}_{E,t}$ entre les données définitives et provisoires est inconnue pour l'année a et est prédite mois par mois à l'aide d'un modèle, ce qui donne une série notée $\widehat{\text{diff}}_{E,t}$. Cette série est agrégée sur l'année a. D'autre part, les données provisoires ne sont pas totalement connues pour l'année a puisque la valeur de décembre est inconnue. Cette valeur est prédite à l'aide d'un autre modèle, ce qui donne une série notée $\widehat{\text{prov}}_{E,t}$. Cette série est aussi agrégée mois par mois sur toute l'année a. L'agrégation des prédictions pour $\widehat{\text{diff}}_{E,t}$ sommée avec l'agrégation de $\widehat{\text{prov}}_{E,t}$ donne la prédiction $\widehat{E}_a$. Dans les sections suivantes, l'erreur de prédiction est examinée en détail. Il s'agit de savoir si celle-ci provient principalement de la composante $\widehat{\text{prov}}_{E,t}$, c'est-à-dire de la prédiction du mois de décembre pour les données provisoires, ou bien de la composante $\widehat{\text{diff}}_{E,t}$, c'est-à-dire de l'évaluation de la différence entre les données définitives et provisoires pour l'année a.

4.2 Performances pour les mariages

Les performances du backtest pour le cas particulier des mariages sont présentées en table 3. La première observation quant aux performances est que l'erreur de prédiction totale est principalement expliquée par l'erreur sur le modèle de la série $\text{diff}_{E,t}$: les erreurs de prédiction de la valeur provisoire du mois de décembre sont généralement faibles par rapport à l'erreur totale.

TABLE 3 Nombre de mariages estimés ($\widehat{E}_a$) et réalisés (E_a) pour les années cibles 2019 à 2024 dans le cadre d'un backtest des flash estimates. $\text{RelError}_{E,a}$ est l'erreur relative de l'équation (3). $\text{Error}_{\text{tot}}$ est la différence $\widehat{E}_a - E_a$. $\text{ErrorProv}_{\text{dec}}$ est la différence entre la valeur prédite et réalisée pour la valeur provisoire de décembre.

évènement	année	$\widehat{E}_a$	E_a	$\text{RelError}_{E,a}$	$\text{Error}_{\text{tot}}$	$\text{ErrorProv}_{\text{dec}}$
mariages	2019	39'701	38'974	1.9%	727	245
mariages	2020	36'102	35'160	2.7%	942	-82
mariages	2021	36'295	36'410	0.3%	-115	68
mariages	2022	37'536	40'938	8.3%	-3'402	133
mariages	2023	42'307	37'763	12.0%	4'544	133
mariages	2024	37'612	36'769	2.3%	843	35

L'année 2019 montre une performance acceptable, avec une erreur relative de 1.9%. En 2020, le modèle de prédiction connaît les outliers liés au COVID-19 dans les données provisoires : des AO en avril, mai et juin 2020, voir table 1. En revanche, les outliers de 2020 sur la série $\text{diff}_{E,t}$ ne sont pas connus. Ces outliers sont deux LS qui sont représentés sur la figure 3 et qui prennent en compte une baisse structurelle de la série en 2020. En conséquence, pour 2020 l'estimation du nombre de mariages $\widehat{E}_a$ surestime la valeur réelle E_a de 2.7%. Les performances pour 2021 sont bonnes, peut-être en raison de l'absence d'outliers cette année-là sur la série $\text{diff}_{E,t}$.

En 2022 et 2023, les performances sont particulièrement mauvaises en raison de la rupture de série de $\text{diff}_{E,t}$ liée à l'autorisation des mariages de personnes du même sexe. Pour 2022, les prédictions qui ne prennent pas du tout en compte cette nouvelle information sous-estiment massivement la valeur finale : il y a 3'402 mariages en plus de l'estimation. Cette valeur est cependant interprétable. En 2022, 749 couples de même sexe se marient et 2'234 couples convertissent leur partenariat enregistré en mariage, soit 2'983 mariages de couples du même sexe. En corrigeant l'erreur de ce chiffre, la sous-estimation n'est plus que de 419 mariages, ce qui correspond à une erreur relative de 1.0%, plus proche des erreurs relatives observées précédemment.

Pour 2023, le modèle REG-ARIMA utilisé montre des limites intéressantes. En effet, à cette date, l'outlier de 2022 sur $\text{diff}_{E,t}$, un LS au mois de juillet, est connu. Le modèle fait implicitement l'hypothèse que la différence $\text{diff}_{E,t}$ va rester, pour toute l'année 2023, sur les nouveaux niveaux très élevés observés depuis juillet 2022. En pratique, ce n'est pas du tout ce qui s'est produit. En 2023, 949 couples de même sexe se marient et 799 convertissent leur partenariat enregistré en mariage, ce qui fait un total de 1'748 mariages de couples de même sexe. Il faut comprendre que le modèle fait implicitement l'hypothèse que le grand nombre (environ 3'000) de mariages de personnes du même sexe enregistré de juillet 2022 à décembre 2022, donc sur une période de 6 mois, va se poursuivre sur toute l'année 2023, c'est-à-dire sur une période deux fois plus longue. Ainsi, le modèle s'attend à obtenir environ 6'000 mariages supplémentaires en 2023 du fait de l'ouverture du mariage aux personnes du même sexe, et ce chiffre de 6'000 se heurte au nombre réalisé qui n'est que de 1'748. Ceci permet d'interpréter la sur-estimation d'environ 4'500 du nombre de mariages en 2023 et les mauvaises performances globales cette année-là.

Pour 2024, les performances commencent à s'améliorer, principalement parce que le nombre de mariages de personnes du même sexe a un comportement qui commence à se stabiliser dans le temps. Sur l'ensemble de l'année 2024, il y a environ 800 mariages de personnes du même sexe et 300 conversions de partenariats enregistrés en mariages, soit au total 1'100 mariages de personnes du même sexe environ : une baisse d'environ 650 mariages par rapport à l'année précédente. La sur-estimation de 843 mariages en 2024, soit une erreur relative de 2.3%, est ainsi toujours due en partie à ce ralentissement ; mais les estimations deviennent tout de même globalement meilleures au fil du temps. Compte tenu de ces interprétations, il est possible d'envisager raisonnablement des erreurs relatives à l'avenir de l'ordre de 2%, c'est-à-dire des sur-estimations ou sous-estimations d'environ 700 mariages à l'avenir.

4.3 Performances pour les divorces

Les performances du backtest pour les divorces sont présentées en table 4. Il est rappelé

Table 4 Nombre de divorces estimés ($\widehat{E}_a$) et réalisés (E_a) pour les années cibles 2019 à 2024 dans le cadre d'un backtest des flash estimates.

évènement	année	$\widehat{E}_a$	E_a	$\text{RelError}_{E,a}$	$\text{Error}_{\text{tot}}$	$\text{ErrorProv}_{\text{dec}}$
divorces	2019	16'484	16'885	2.4%	-401	-43
divorces	2020	16'680	16'210	2.9%	470	-96
divorces	2021	17'497	17'159	2.0%	338	292
divorces	2022	16'255	16'201	0.3%	54	174
divorces	2023	15'341	15'562	1.4%	-221	-37
divorces	2024	15'994	16'123	0.8%	-129	-14

qu'avec la modélisation actuelle, les deux modèles associés aux divorces ne contiennent aucun outlier, ce qui simplifie les interprétations par rapport à ce qui a été observé pour les mariages. Les performances observées montrent des erreurs relatives globalement inférieures à 2.5%, à l'exception de 2020 où les divorces sont sur-estimés de 2.9%. Il est possible d'observer cependant que l'erreur de prévision pour le seul mois de décembre sur les données provisoires est parfois assez importante, avec par exemple une sur-estimation de 292 divorces pour décembre 2021 ou bien 174 pour décembre 2022. Il existe peut-être un potentiel d'amélioration du modèle quant à la modélisation des effets calendrier sur la série provisoire des divorces, et en particulier sur les dates légales de fermetures des tribunaux à Pâques et au mois de décembre, c'est-à-dire justement le mois où des prédictions sont effectuées.

4.4 Performances pour les naissances

Les performances du backtest pour les naissances sont données en table 5. Tout comme pour

Table 5 Nombre de naissances estimées ($\widehat{E}_a$) et réalisés (E_a) pour les années cibles 2019 à 2024 dans le cadre d'un backtest des flash estimates.

évènement	année	$\widehat{E}_a$	E_a	RelError$_{E,a}$	Error$_{\text{tot}}$	ErrorProv$_{\text{dec}}$
naissances	2019	87'152	86'172	1.1%	980	155
naissances	2020	85'097	85'914	1.0%	-817	-55
naissances	2021	89'939	89'644	0.3%	295	136
naissances	2022	82'238	82'371	0.2%	-133	-44
naissances	2023	80'509	80'024	0.6%	485	73
naissances	2024	79'178	78'256	1.2%	922	90

les divorces, les modèles associés aux naissances ne contiennent pas d'outliers. Les performances observées sont les meilleures des 4 séries MDND étudiées, avec des erreurs relatives de l'ordre de 1%. Ces erreurs ne semblent pas être liées à de mauvaises estimations de la valeur provisoire de décembre : l'erreur totale s'explique presque entièrement par l'erreur de prévision du modèle sur diff$_{E,t}$.

4.5 Performances pour les décès

Les performances du backtest pour les décès sont données en table 6. Les résultats observés

Table 6 Nombre de décès estimés ($\widehat{E}_a$) et réalisés (E_a) pour les années cibles 2019 à 2024 dans le cadre d'un backtest des flash estimates.

évènement	année	$\widehat{E}_a$	E_a	RelError$_{E,a}$	Error$_{\text{tot}}$	ErrorProv$_{\text{dec}}$
décès	2019	68'010	67'780	0.3%	230	226
décès	2020	72'402	76'195	5.0%	-3'793	-3'403
décès	2021	70'099	71'192	1.5%	-1'093	-1'124
décès	2022	73'536	74'425	1.2%	-889	-928
décès	2023	72'564	71'822	1.0%	742	443
décès	2024	72'330	71'942	0.5%	388	405

sont contrastés. En 2019, les performances du modèle sont bonnes avec une erreur relative

de 0.3%. L'année 2020 montre des performances particulièrement mauvaises, avec une sous-estimation du nombre de décès de presque 3'800. Pour interpréter ce chiffre, il faut rappeler les conditions du backtest pour l'année 2020. Les données provisoires des décès jusqu'au mois de novembre sont connues, et en particulier les AO en avril 2020 et novembre 2020 liés à la première et seconde vague de la pandémie. Par contre, la valeur provisoire de décembre 2020 n'est pas connue et est prédite. Comme le modèle considère que le mois de novembre 2020 est un AO, c'est-à-dire une observation ponctuelle qui est un outlier, celui-ci fait implicitement l'hypothèse que la situation se normalise en décembre et que le nombre de décès revient à des valeurs plus habituelles pour un mois de décembre. Ainsi, en 2020, le modèle des données provisoires sous-estime le nombre de décès en décembre d'environ 3'400, ce qui explique l'essentiel de l'erreur effectuée sur 2020. Le backtest pour l'année 2020 n'intègre pas l'information que la seconde vague de la pandémie a continué de frapper en décembre 2020, ce qui explique l'erreur relative élevée : 5%.

En 2021, le modèle continue de sous-estimer le nombre de décès, mais dans des proportions moindres avec une erreur relative de 1.5%. L'erreur est une nouvelle fois entièrement liée à une sous-estimation du nombre de décès en décembre. La même situation se reproduit en 2022. En 2023 et 2024, les performances s'améliorent et le modèle des données provisoires cesse de donner une sous-estimation. Globalement, il est clair que pour chaque année cible étudiée, l'erreur totale de prédiction est bien expliquée par l'erreur de prédiction de la valeur provisoire de décembre. Toute tentative d'amélioration des performances devrait donc se focaliser sur le modèle des données provisoires et non sur le modèle pour $\text{diff}_{E,t}$.

5 Travail de production des flash estimates

Pour la production des flash estimates sur une année a donnée, des scripts en langage R sont mis en place de sorte à assurer une production efficace et interprétable en début d'année $a+1$. Un script est créé pour chaque évènement E étudié ici, soit 4 scripts au total. Les étapes d'un script sont les suivantes :

1. Chargement des données définitives et provisoires, avec imputations, jusqu'à la fin de l'année $a-1$.
2. Saisie des 11 valeurs provisoires allant des mois de janvier à novembre de l'année a.
3. Prédiction, à l'aide du modèle des données provisoires, de la valeur provisoire manquante de décembre de l'année a.
4. Prédiction, à l'aide du modèle sur $\text{diff}_{E,t}$, de la différence entre les données définitives et les données provisoires sur l'ensemble de l'année a. Cette prédiction se base sur les données jusqu'à la fin de l'année $a-1$ et ne dépend pas de la prédiction de l'étape précédente sur les données provisoires, ni même des valeurs provisoires de janvier à novembre de l'année a.
5. Création de la série de données provisoires "complétée" $\widehat{\text{prov}}_{E,t}$, et calcul de la somme (pour l'année a) de cette série complétée avec la prédiction $\widehat{\text{diff}}_{E,t}$ effectuée sur la différence "définitif-provisoire". Le résultat est une prédiction des données définitives mois par mois pour l'ensemble de l'année a, soit 12 valeurs.
6. Calcul de la somme des 12 prédictions mensuelles des données définitives pour produire $\widehat{E}_a$. Graphiques et interprétations.

L'étape 3 décrite ci-dessus demande quelques précisions. Il est possible que certaines des 11 valeurs provisoires, saisies lors de l'étape 2, soient considérées comme des outliers dans le

modèle de l'étape 3. Le modèle de l'étape 3 est ainsi établi en fixant les outliers jusqu'à la fin de l'année $a-1$, et en autorisant une détection automatique d'outliers supplémentaires uniquement sur la période qui couvre l'année a. Lorsque la prédiction de l'étape 3 est produite, il est ainsi important de contrôler si de nouveaux outliers ont été détectés et le cas échéant de faire le travail d'interprétation et de discussion avec la section DEM. Un autre point à vérifier avant l'étape 3 est la présence éventuelle de données provisoires incomplètes, voire de données manquantes qui demanderaient une imputation directement lors de l'étape 2. Ce point n'est pas simple à traiter car la stratégie d'imputation décrite dans ce rapport, qui impute la série $\text{diff}_{E,t}$ pour en déduire une imputation sur les données provisoires, ne peut pas être directement répliquée ici. En effet, les données définitives de l'année a sont inconnues, ce qui ne permet plus de calculer un lien simple entre les données provisoires et la série $\text{diff}_{E,t}$. Pour le moment, en cas de donnée manquante pour l'année a ou s'il est soupçonné qu'une valeur provisoire est incomplète, il est décidé d'effectuer l'imputation sur la base du modèle des données provisoires, et de réviser cette imputation plus tard, une fois que les données définitives sont disponibles.

Un exemple de prédiction, obtenu pour les mariages de l'année 2025 est donné en figure 10. Le nombre de mariages prédits pour toute l'année 2025 est de 37'054. Au moment où cette

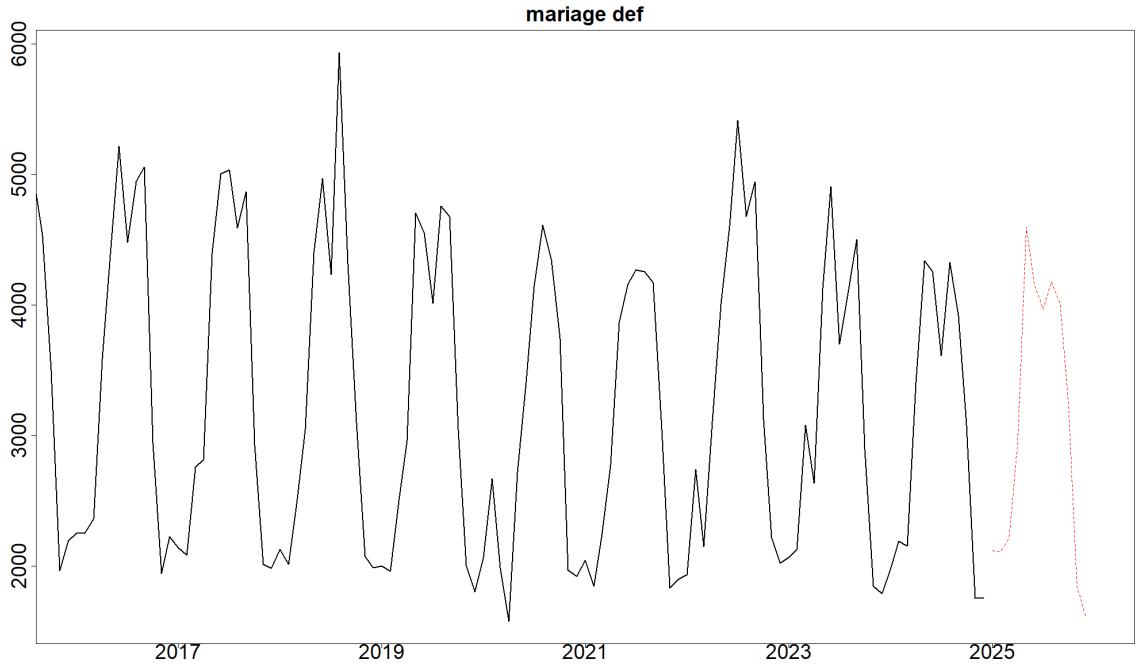


FIGURE 10 Données définitives pour les mariages jusqu'à fin 2024 (courbe noire). Prédiction pour toute l'année 2025 à l'aide des modèles et de la méthodologie présentée.

estimation a été produite, il était anticipé que l'estimation était peut-être sur-estimée en raison d'une possible sur-estimation du nombre de mariages de couples du même sexe. L'estimation peut être confrontée à la valeur définitive, qui a été publiée entre temps au mois de juin 2026 : 35'913 mariages sur toute l'année 2025. L'estimation produite surestime le nombre de mariages de 1'141 unités, soit une erreur relative de 3.1%. L'absence de Level Shift négatif en 2025 dans la modélisation des mariages fait ainsi perdurer la surestimation qui était déjà présente pour l'année 2024, voir table 3, d'autant plus que le nombre de mariages de personnes du même sexe continue de légèrement diminuer en 2025. Pour les prédictions des mariages en 2026, un Level Shift sera ajouté au modèle de la série $\text{diff}_{E,t}$ ce qui devrait résoudre le problème persistant de surestimation.

Les résultats pour les naissances et les divorces ne posent pas de problèmes particuliers. La prédiction pour les naissances en 2025 est de 78'781 naissances. Le nombre de naissances en 2025 est de 78'153. L'estimation a ainsi une erreur relative de 0.8%. La prédiction pour les divorces est de 16'184. Le nombre de divorces en 2025 est plus faible qu'attendu à 15'765, ce qui fait une erreur relative de 2.7%. Pour les décès, un problème d'enregistrement dans les données provisoires est présent. Tout comme les données de novembre et décembre 2024, qui sont déjà imputées, il semble que les données provisoires de janvier à avril 2025 pourraient également être incomplètes. Un examen détaillé de ces données montre des valeurs plutôt conformes aux prédictions de la série provisoire pour janvier et février 2025, mais nettement plus basses que les prédictions pour mars et avril 2025. Il est donc décidé pour le moment de ne pas imputer/modifier les valeurs de janvier et février 2025 et de remplacer les valeurs communiquées par DEM pour mars et avril 2025 par une imputation. L'imputation se fait à l'aide du modèle sur les données provisoires, comme expliqué plus haut dans cette section. Sur cette base, le nombre de décès prévus en 2025 est de 71'430. Les prédictions sont représentées en figure 11. Le nombre

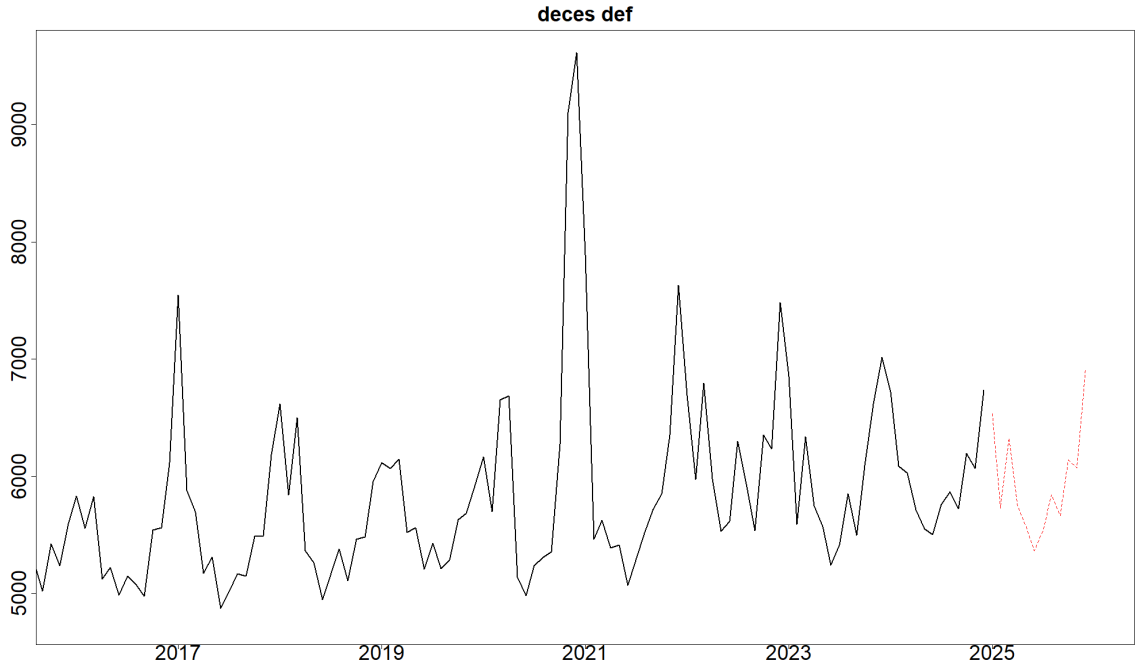


FIGURE 11 Données définitives pour les décès jusqu'à fin 2024 (courbe noire). Prédiction pour toute l'année 2025 à l'aide des modèles et de la méthodologie présentée.

de décès réalisé en 2025 est de 71'904. L'erreur relative associée à l'estimation produite par le modèle est ainsi de 0.7%.

6 Conclusions et perspectives

L'utilisation d'une stratégie et d'une modélisation REG-ARIMA adaptée permet de produire, dès le début d'une année donnée, des flash estimates du nombre de mariages, divorces, naissances et décès au cours de l'année écoulée. Les publications définitives étant généralement effectuées au mois de juin, les utilisateurs peuvent ainsi bénéficier d'une estimation des résultats près de 4 mois plus tôt. Les résultats obtenus montrent que la production de flash estimates est techniquement réalisable et permet d'obtenir, dans la majorité des situations étudiées, des estimations fiables. L'erreur de prédiction observée en backtest est de l'ordre de 2% de la valeur définitive pour les naissances, décès ou divorces et légèrement supérieure à 2% pour les mariages. Des évènements particuliers peuvent toutefois venir affecter les performances, comme le montrent les backtests effectués. Ceci est aussi valable en cas de problèmes dans l'intégrité des données provisoires qui pourraient fausser les estimations.

Les performances obtenues sont globalement encourageantes, d'autant plus que la modélisation présente encore des potentiels d'amélioration. Pour les mariages, celles-ci pourraient porter sur la gestion des dates insolites et d'une modélisation plus adaptée pour la prédiction des mariages de couples du même sexe, autorisés depuis l'été 2022. Pour les divorces, ce sont les aspects liés au calendrier qui pourraient être approfondis, avec notamment une meilleure prise en compte des dates d'ouverture des tribunaux dédiés. Pour les décès, le potentiel pourrait résider sur une prédiction optimisée du nombre de décès en décembre. Pour les naissances, les performances sont les meilleures présentées ici. Les améliorations pourraient porter sur le modèle ARIMA utilisé pour les prédictions, et peut-être une meilleure compréhension de la tendance, avec des naissances qui baissent ces dernières années.

Les résultats produits par les flash estimates devraient être publiés en tant que statistique expérimentale. En publiant ces flash estimates, l'OFS rejoint d'autres pays européens comme la France pour qui l'INSEE publie à la mi-janvier des statistiques pour l'année révolue [5]. Ces estimations ne doivent toutefois pas être considérées comme un substitut aux statistiques définitives publiées par l'OFS, mais comme une indication précoce de leur valeur probable.

5. https://www.insee.fr/fr/statistiques/8719824

Bibliographie

Box, G.E.P. & Jenkins, G.M. & Reinsel, G.C. & Jung, G.M. (2015). Time Series Analysis : Forecasting and Control. *John Wiley & Sons*

Brockwell, P.J. & Davis, R.A. (2009). Time Series : Theory and Methods. *Springer science & business media*

Chevalier, C. (2023). Rapport de méthodes : Méthodes de correction des effets calendrier en usage à l'OFS en 2023. Rapport de méthodes, Office Fédéral de la Statistique OFS. URL https://www.bfs.admin.ch/bfs/fr/home/services/recherche/rapports-methodologiques.assetdetail.26565928.html.

R Core Team (2021). R : A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Sax, C. & Eddelbuettel, D. (2018). Seasonal Adjustment by X-13ARIMA-SEATS in R. *Journal of Statistical Software* **87**, 1–17.

U.S. Census Bureau (2025). Official X-13-ARIMA-SEATS manual. Tech. rep., U.S. Census Bureau.
URL https://www2.census.gov/software/x-13arima-seats/x13as/unix-linux/documentation-/docx13ashtml.pdf.

Annexe

A Exemple d'analyse de révision pour le modèle des données provisoires des mariages

En section 3.3.4, un modèle ARIMA(1,1,1)(0,1,1) est ajusté pour les données provisoires des mariages. L'analyse des révisions consiste à observer comment les paramètres du modèle varient si le modèle est ajusté à l'aide d'une partie des données et non avec les données complètes. Dans cet exemple, X13-ARIMA-SEATS estime les paramètres du modèle en utilisant toutes les données jusqu'à janvier 2019, février 2019, ..., décembre 2024, ce qui correspond à 72 ajustements différents, et donc 72 jeux de paramètres estimés. Chaque paramètre est donc estimé 72 fois en faisant varier la date des dernières données utilisées pour l'ajustement, de janvier 2019 à décembre 2024. Les boîtes à moustaches de la figure 12 rendent compte de la stabilité des estimations pour chaque paramètre du modèle. Les paramètres estimés dans les analyses de révision sont les paramètres de régression correspondant aux effets calendrier et les paramètres du modèle ARIMA, à l'exception du paramètre σ^2, la variance du bruit blanc. X13-ARIMA-SEATS ne permet pas de faire d'analyse de révision sur les paramètres associés aux outliers.

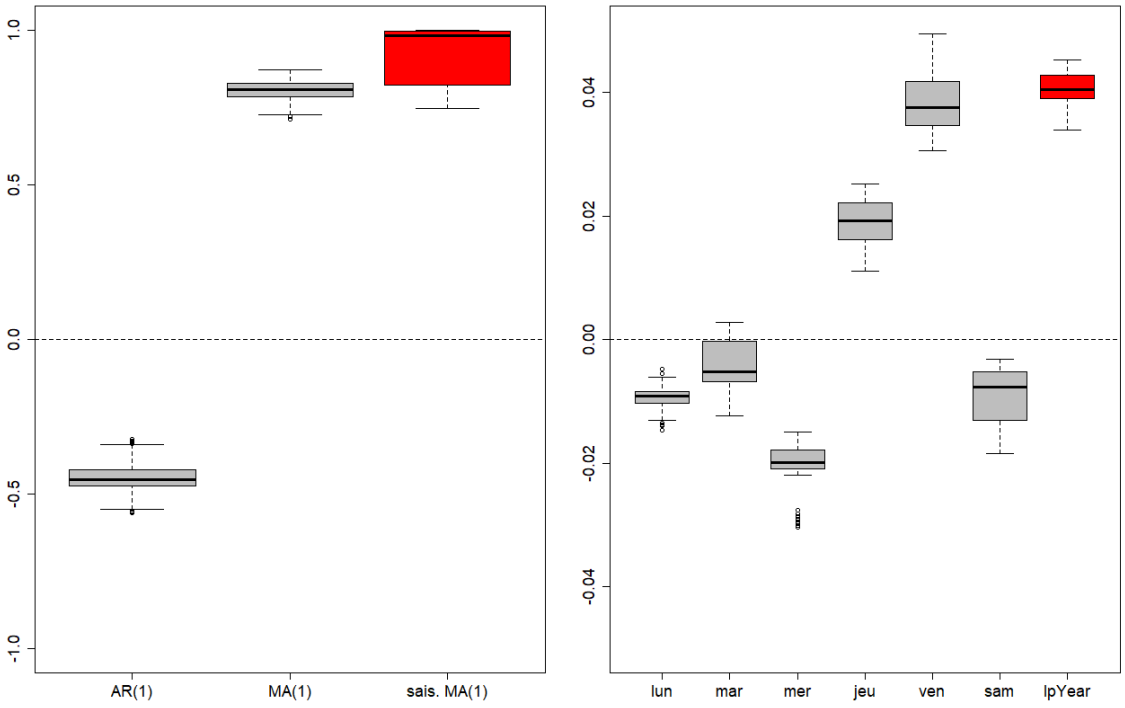


FIGURE 12 Analyse de révision du modèle des données provisoires des mariages. L'ajustement des modèles se fait sur la base de séries dont la dernière valeur mensuelle utilisée va de janvier 2019 à décembre 2024, soit 72 ajustements au total. Les boîtes à moustaches sont associées aux paramètres du modèle ARIMA (graphique de gauche) et aux paramètres liés aux effets calendrier (graphique de droite). Sur le graphique de gauche, AR(1), MA(1) et sais.MA(1) désignent respectivement les paramètres ϕ_1, θ_1 et Θ_1, voir équation (2). Sur le graphique de droite, lun, mar, ..., sam désignent les paramètres associés aux effets des lundis, mardis, ..., samedis non-fériés. Lpyear est le paramètre associé aux effets des années bissextiles.

Rapports de méthodes de la section méthodes statistiques de l'OFS
Methodenberichte der Sektion Statistische Methoden des BFS
Methodology reports published by the FSO's Statistical Methods Section

Chevalier, C., Meskaldji, D.-E. (2026). Statistique du mouvement naturel de la population. Flash estimates. Numéro OFS : 338-0096

Meskaldji, D.-E., Mariétan, R. (2026). Applications du bootstrap à l'OFS. Estimation de variance à l'aide de méthodes de rééchantillonnage. Numéro OFS : 338-0095

Chevalier, C. (2025). Statistique sur les télécommunications. Plan de sondage. Numéro OFS : 338-0094

Potterat, J. (2025). Erhebung leichte Nutzfahrzeuge (LWE) 2023. Stichprobenplan, Gewichtung und Schätzverfahren. BFS-Nummer : 338-0093

Chevalier, C. (2025). Statistique de la valeur ajoutée 2021. Révision de la pondération. Numéro OFS : 338-0092

Bianchi, B., Leuenberger, M., Massiani, A., Panchard, C. (2025). Enquête PIAAC 2022-2023. Plan de sondage, pondération, estimation de variance et analyse du biais dû à la non-réponse. Numéro OFS : 338-0091

Meskaldji, D.-E., Panchard, C., Qualité, L., Saliba, J. (2024). Enquêtes thématiques 2018-2023. Plans de sondage, tirages, pondération et estimation de variance. Numéro OFS : 338-0090

Chevalier, C. (2024). Statistique de la valeur ajoutée 2021. Révision du plan de sondage. Numéro OFS : 338-0089

Nedyalkova, D., Assoulin, D. (2024). Swiss structural business statistics (STATENT) 2015-2021. Estimation and precision of the full-time equivalents. FSO number : 338-0088

Qualité, L. (2023). Relevé structurel de la population suisse 2018. Révision de la pondération et sélections répétées. Numéro OFS : 338-0087

Potterat, J., Assoulin, D. (2023). Forschung und Entwicklung 2021. Stichprobenplan, Gewichtung und Schätzverfahren. BFS-Nummer : 338-0086

Chevalier, C. (2023). Méthodes de correction des effets calendrier en usage à l'OFS en 2023. Construction, choix et application des régresseurs mis à disposition. Numéro OFS : 338-0085

Mariétan, R. (2022). Comportement en matière de voyages 2020. Révision de la pondération. Numéro OFS : 338-0084

Massiani, A. (2021). Enquête suisse sur la population active 2015-2020 : Révision de la pondération et de l'estimation de la variance. Numéro OFS : 338-0083

Chevalier, C., Qualité, L. (2021). Enquête suisse sur la structure des salaires 2018. Révision de la pondération. Numéro OFS : 338-0082

Qualité, L., Potterat, J. (2020). Enquête suisse sur la structure des salaires 2018. Révision du plan de sondage. Numéro OFS : 338-0081

Potterat, J., Qualité, L., Assoulin, D. (2019). Strukturerhebung der eidgenössischen Volkszählung. Stichprobenplan und -ziehung, Gewichtung, Schätzverfahren und Pooling 2010-2018. BFS-Nummer : 338-0080

Potterat, J., Assoulin, D., Nicoletti, J.-M. (2019). Beschäftigungsstatistik BESTA. Revision 2015 : Stichprobenrahmen, Stichprobenplan und Hochrechnung. BFS-Nummer : 338-0079

Massiani, A. (2017). Estimation de la couverture du nouveau recensement en Suisse en 2012. Numéro OFS : 338-0078

Nedyalkova, D., Assoulin, D. (2017). Construction of full-time equivalents for the Swiss structural business statistics. FSO number : 338-0077

Ferster, M. (2016). Energieverbrauchsstatistik EVS2014 - Stichprobe, Hochrechnung und Vergleichbarkeit mit der EVS2013. BFS-Nummer : 338-0076

Panchard, C. (2014). Enquête sur la participation aux activités culturelles 2008. Plan de sondage et estimations. BFS-Nummer : 338-0073

Assoulin, D. (2014). Wertschöpfungsstatistik. Revision 2009 : Statistische Datenaufbereitung und Hochrechnung. BFS-Nummer : 338-0072

Wilhelm, M. (2014). Echantillonnage boule de neige. La méthode de sondage déterminé par les répondants. Numéro OFS : 338-0071

Assoulin, D. (2013). Wertschöpfungsstatistik. Revision 2009 : Stichprobenrahmen und Stichprobenplan. BFS-Nummer : 338-0070

Ferster, M. (2013). EVS I - Energieverbrauchsstatistik 2002 bis 2007 : Stichprobenplan und Hochrechnung. BFS-Nummer : 338-0069

Assoulin, D. (2013). Zusatzerhebung für die landwirtschaftliche Betriebszählung 2010 : Stichprobenplan und Hochrechnung. BFS-Nummer : 338-0068

Potterat, J. (2012). Use of conversion keys for NOGA2002-2008. FSO number : 338-0067-05

Andrade, B., Salamin P.-A. (2012). Enquête sur la situation sociale et économique des étudiant-e-s des hautes écoles suisses 2009. Cadre de sondage, plan d'échantillonnage et méthodes d'estimation. Numéro OFS : 338-0066

Potterat, J., Panchard, C., Kilchmann, D. (2012). Umweltschutzausgaben der Unternehmen 2009 (UWSA2009). Stichprobenplan, Einsetzungen, Gewichtung und Schätzverfahren. BFS-Nummer : 338-0065

Potterat, J. (2012). Benutzung der Umsteigeschlüssel NOGA 2002-2008. BFS-Nummer : 338-0064

Potterat, J. (2011). Kosten und Nutzen der Berufsbildung aus Sicht der Betriebe im Jahr 2009 (KNBB09). Stichprobenplan, Gewichtung und Schätzverfahren. BFS-Nummer : 338-0063

Kilchmann, D., Potterat, J., Genoud, S. (2011). Gütertransporterhebung 2008. Stichprobenplan, Datenaufbereitung, Gewichtung und Schätzverfahren. BFS-Nummer : 338-0062

Graf, E. (2010). Enquête suisse sur la santé 2007. Plan d'échantillonnage, pondérations et analyses pondérées des données. Numéro OFS : 338-0061

Eichenberger P., Hulliger B., Potterat J. (2010). Describing the Anticipated Accuracy of the Swiss Population Survey. FSO number : 338-0060

Graf, E. (2010). Étude empirique de l'attrition du Panel Suisse de Ménages : vers une caractérisation du profil du non-répondant. Numéro OFS : 338-0059

Graf, E. (2009). Weightings of the Swiss Household Panel : SHP_I wave 9, SHP_II wave 4, SHP_I et SHP_II combined. FSO number : 338-0058

Graf, E. (2009). Pondérations du Panel Suisse de Ménages : PSM_I vague 9, PSM_II vague 4, PSM_I et PSM_II combinés. Numéro OFS : 338-0057-05

Qualité, L., Tillé, Y. (2009). Estimation de la précision d'évolutions dans l'enquête sur la valeur ajoutée. Numéro OFS : 338-0056

Renaud, A., Panchard, C. et Potterat, J. (2008). Statistique de l'emploi. Révision 2007 : méthodes d'estimation. Numéro OFS : 338-0055

Graf, E. (2008). Pondérations du PSM. PSM_I vague 8, PSM_II vague 3, PSM_I et PSM_II combinés. Numéro OFS : 338-0054

Andrade, B., Graf, M. (2008). Enquête suisse sur la structure des salaires 2006. Aspects méthodologiques du modèle des salaires "Salarium". Numéro OFS : 338-0053

Renaud, A. (2008). Statistique de l'emploi. Révision 2007 : cadre de sondage et échantillonage. Numéro OFS : 338-0052

Graf, E. (2008). Pondérations du SILC pilote. SILC_I vague 2, SILC_II vague 1, SILC_I et SILC_II combinés. Numéro OFS : 338-0051

Kilchmann, D. (2008). Statistik der sozialmedizinischen Institutionen 1999-2004 und Krankenhausstatistik 1999-2002. Einsetzungen für fehlende Daten. BFS-Nummer : 338-0050

Renaud, A. (2008). Technologies de l'information et de la communication. Estimations sur la base de la statistique de la valeur ajoutée. Numéro OFS : 338-0049

Assoulin, D. (2007). Wertschöpfungsstatistik. Einsetzungsversuche für fehlende Antworten grosser Unternehmen. BFS-Nummer : 338-0048

Kilchmann, D. (2007). Beherbergungsstatistik Campingplätze. Stichprobenrahmen und Schätzverfahren 2005/06. BFS-Nummer : 338-0047

Gabler, S., Häder, S. (2007). Haushalts- und Personenerhebungen. Machbarkeit von Random Digit Dialing in der Schweiz. BFS-Nummer : 338-0046

Ferrez, J., Graf, M. (2007). Enquête suisse sur la structure des salaires. Programmes R pour l'intervalle de confiance de la médiane. Numéro OFS : 338-0045

Renaud, A. (2007). Harmonisation de la scolarité obligatoire en Suisse (HarmoS). Design général de l'enquête et échantillon des écoles. Numéro OFS : 338-0044

Potterat, J. (2007). Betriebszählung 2005. Statistische Methoden zur Schätzung der provisorischen Ergebnisse. BFS-Nummer : 338-0043

Hulliger, B. (2006). Umweltschutzausgaben der Unternehmen 2003, Stichprobenplan, Datenaufbereitung und Schätzverfahren. BFS-Nummer : 338-0042

Renfer, J.-P. (2006). Enquête sur les chiffres d'affaires du commerce de détail. Plan d'échantillonnage et méthodes d'estimation. Numéro OFS : 338-0041

Salamin, P.-A. (2006). Statistique de l'aide sociale dans le domaine de l'asile. Plan de sondage et extrapolations pour l'enquête pilote 2005. Numéro OFS : 338-0040

Renaud, A. (2006). Statistique suisse des bénéficiaires de l'aide sociale. Pondération des communes 2004. Numéro OFS : 338-0039

Graf, M. (2006). Swiss Earnings Structure Survey 2002-2004. Compositional data in a stratified two-stage sample : Analysis and precision assessment of wage components. FSO number : 338-0038

Potterat, J. (2006). Pensionskassenstatistik 2004. Statistische Methoden zur Schätzung der provisorischen Ergebnisse. BFS-Nummer : 338-0037

Potterat, J. (2006). Kosten und Nutzen der Berufsbildung aus Sicht der Betriebe im Jahr 2004. Stichprobenplan, Gewichtung und Schätzverfahren. BFS-Nummer : 338-0036

Kilchmann, D. (2006). Vierteljährliche Wohnbaustatistik. Stichprobenplan, statistische Datenaufarbeitung und Schätzverfahren 2005. BFS-Nummer : 338-0035

Kilchmann, D. (2006). Erhebung über Forschung und Entwicklung in der schweizerischen Privatwirtschaft 2004. Bereinigung der Stichprobe, Ersatz fehlender Werte und Schätzverfahren. BFS-Nummer : 338-0034

Kilchmann, D., Eichenberger, P., Potterat, J. (2005). Volkszählung 2000. Statistische Einsetzungsverfahren Band 2. BFS-Nummer : 338-0033

Kilchmann, D., Eichenberger, P., Potterat, J. (2005). Volkszählung 2000. Statistische Einsetzungsverfahren Band 1. BFS-Nummer : 338-0032

Graf, M., Matei, A. (2005). Enquête suisse sur la structure des salaires 2002. La précision du salaire brut standardisé médian. Numéro OFS : 338-0031

Graf, E., Renfer, J.-P. (2005). Enquête suisse sur la santé 2002. Plan d'échantillonnage, pondération et estimation de la précision. Numéro OFS : 338-0030

Potterat, J. (2005). Mietpreis-Strukturerhebung 2003. Gewichtung und Schätzverfahren. BFS-Nummer : 338-0029

Potterat, J. (2005). Landwirtschaftliche Betriebszählung 2003. Schätzverfahren für die Zusatzerhebung. BFS-Nummer : 338-0028

Renaud, A. (2004). Coverage estimation for the Swiss population census 2000. Estimation methodology and results. FSO number : 338-0027

Kilchmann, D. (2004). Revision des Schweizerischen Lohnindex. Schätzmethoden der Lohnindices und deren Varianzschätzer. BFS-Nummer : 338-0026

Graf, M. (2004). Enquête suisse sur la structure des salaires 2002. Plan d'échantillonnage et extrapolation pour le secteur privé. Numéro OFS : 338-0025

Renaud, A. (2004). Analyse de données d'enquêtes. Quelques méthodes et illustration avec des données de l'OFS. Numéro OFS : 338-0024

Renaud, A., Potterat, J. (2004). Estimation de la couverture du recensement de la population de l'an 2000. Echantillon pour l'estimation de la sous-couverture (P-sample) et qualité du cadre de sondage des bâtiments. Numéro OFS : 338-0023

Graf, M. (2004). Fusion de données. Etude de faisabilité. Numéro OFS : 338-0022

Potterat, J. (2003). Mietpreis-Strukturerhebung 2003. Entwicklung des Stichprobenplans und Ziehung der Stichprobe. BFS-Nummer : 338-0021

Potterat, J. (2003). Landwirtschaftliche Betriebszählung 2003. Stichprobenplan der Zusatzerhebung. BFS-Nummer : 338-0020.

Renaud, A. (2003). Estimation de la couverture du recensement de la population de l'an 2000. Echantillon pour l'estimation de la sur-couverture (E-sample). Numéro OFS : 338-0019

Hulliger, B. (2003). Bereinigung der Stichprobe, Ersatz fehlender Werte und Schätzverfahren. Erhebung über F+E in der schweizerischen Privatwirtschaft 2000. BFS-Nummer : 338-0018

Renfer, J.-P. (2003). Enquête 2000 sur la recherche et le développement dans l'économie privée en Suisse. Plan d'échantillonnage. Numéro OFS : 338-0017

Potterat, J. (2003). Kosten und Nutzen der Berufsbildung aus Sicht der Betriebe. Schätzverfahren. BFS-Nummer : 338-0016

Graf, M., Matei, A. (2003). Stratégie de choix des modèles de désaisonnalisation. Application aux séries de l'emploi total. Numéro OFS : 338-0015

Potterat, J., Salamin, P.A. (2002). Betriebszählung 2001. Methoden für die Datenbereinigung. BFS-Nummer : 338-0014

Renaud, A. (2002). Programme international pour le suivi des acquis des é lèves (PISA). Plans d'échantillonnage pour PISA 2000 en Suisse. Numéro OFS : 338-0013

Renfer, J.-P. (2002). Enquête 2001 sur les coûts et l'utilité de la formation des apprentis du point de vue des établissements. Plan d' échantillonnage. Numéro OFS : 338-0012

Potterat, J., Salamin, P.A. (2002). Betriebszählung 2001. Stichprobenplan und Schätzverfahren für die provisorischen Ergebnisse. BFS-Nummer : 338-0011

Graf, M. (2002). Enquête suisse sur la structure des salaires 2000. Plan d'échantillonnage, pondération et méthode d'estimation pour le secteur privé. Numéro OFS : 338-0010

Renaud, A., Eichenberger P. (2002). Estimation de la couverture du recensement de la population de l'an 2000. Procédure d'enquête et plan d'échantillonnage de l'enquête de couverture. Numéro OFS : 338-0009

Kilchmann, D., Hulliger, B. (2002). Stichprobenplan für die Obstbaumzählung 2001. BFS-Nummer : 338-0008

Graf, M. (2002). Passage du concept établissement au concept entreprise. Numéro OFS : 338-0007

Salamin, P.A. (2001). La technique de la double enquête pour la statistique du transport routier de marchandise. Numéro OFS : 338-0006

Peters, R., Renfer, J.-P. et Hulliger, B. (2001). Statistique de la valeur ajoutée 1997-1998. Procédure d'extrapolation des données. Numéro OFS : 338-0005

La statistique du mouvement naturel de la population est publiée chaque année par l'Office Fédéral de la Statistique. Le nombre total de mariages, divorces, naissances et décès au cours de l'année écoulée est publié au mois de juin de l'année suivante. L'objectif de ce travail est d'étudier la possibilité et la pertinence de publier des «flash estimates» plusieurs mois avant la publication finale. Les flash estimates sont effectuées à partir des données provisoires et se basent sur une modélisation REG-ARIMA de séries temporelles. Celle-ci n'offre cependant pas de garantie théorique sur la précision des estimations. En particulier des évènements importants peuvent causer des changements dans le comportement des séries et ainsi affecter la précision des résultats, du fait d'une violation de l'hypothèse de stationnarité sous-jacente aux modèles. Ceci est souligné dans le rapport à travers des exemples. Le choix de paramètres des modèles est décrit en détail et une analyse des performances, sous la forme d'un backtest, est fournie. Les backtests montrent que l'utilisation de données provisoires, disponibles avant les données définitives, et d'une modélisation adaptée permet d'obtenir des écarts de l'ordre de 2% entre les flash estimates et les chiffres définitifs.

En ligne
www.statistique.admin.ch/338-0096

Numéro OFS
338-0096

Distribution
www.publicationsfederales.admin.ch
No d'art. 317.098.F

ISBN
978-3-303-00788-4