MÉTHODE POUR ÉVITER LA MENTION DE DATES EXACTES DANS LES SÉRIES DE DONNÉES QUI SONT MISES À LA DISPOSITION DES CHERCHEURS

Contexte

Lors de l'examen de demandes relatives à la communication de séries de données à des chercheurs, le Comité de sécurité de l'information est régulièrement confronté à la demande de dates exactes. Ces dates doivent être considérées comme un « quasi-identifiant », ce qui rend réel le risque d'identification de l'intéressé même si les identifiants sont pseudonymisés ou ne sont pas présents dans la série de données.

Dans la mesure du possible, la demande d'informations adressée au Comité de sécurité de l'information sera adaptée de sorte à ne plus demander des dates mais plutôt des périodes. Cependant, ceci ne sera pas toujours une solution faisable pour le chercheur dans la mesure où ce dernier utilise souvent des dates pour calculer des durées.

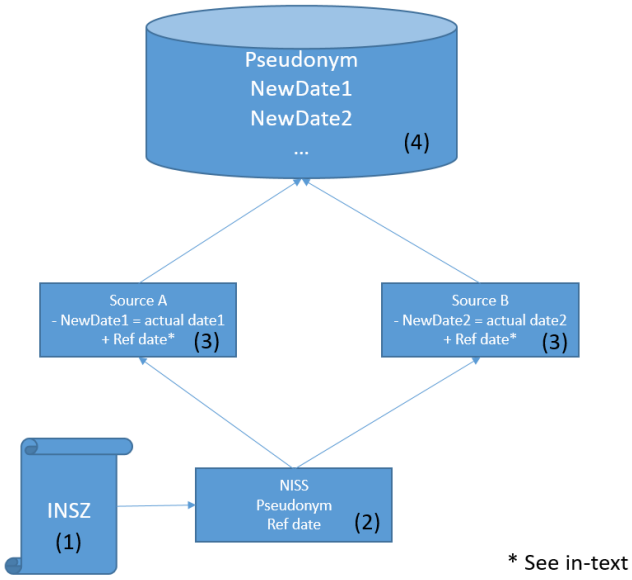
La complexité de la réponse augmente encore lorsque les dates proviennent de deux sources authentiques différentes. Dans ce cas, un TTP devrait interpréter l'information et réaliser des calculs, mais ceci ne relève pas nécessairement de la mission du TTP et limiterait fortement la liberté du chercheur.

Afin de fournir une réponse adéquate à ce problème, une proposition a été élaborée. Dans cette proposition, les dates exactes ne seraient plus communiquées par une source authentique, mais les dates seraient recalculées par rapport à une date de référence fixe, déterminée de manière aléatoire par intéressé dont les données figurent dans la série de données. Dès lors que la date de référence n'apparaît pas dans la série de données finale, il ne sera plus possible de recalculer les dates exactes à partir de cette série de données.

Ceci permet d'offrir une solution simple et rentable pour la plupart des demandes qui requièrent des dates exactes.

Proposition et solution technique

La proposition repose sur le principe suivant :



0. Le domaine temporel de l'étude est déterminé et le nombre de jours (MaxDays) est calculé.
1. Une liste de NISS est établie pour déterminer l'échantillon de l'étude.
2. Pour chacun de ces NISS, un TPP:
    a. créera un pseudonyme et
    b. et déterminera un nombre naturel aléatoire[1] (Ref date) inférieur au nombre de jours du domaine temporel.
3. Chaque source authentique recalculera ses NewDates à l'aide de la formule suivante[2]:
    a. NewDate = (date de début domaine temporel) + ((date réelle - date de début domaine temporel) + (Ref date)) mod (MaxDays)
4. Pour le calcul des durées, le chercheur doit appliquer la formule suivante :
    a. Duration = (NewDate2 – NewDate1 + MaxDays) mod (MaxDays)

Détermination du domaine temporel

La première étape de cette solution consiste à déterminer un domaine temporel qui sera utilisé pour le décalage aléatoire des dates par intéressé.

De manière générale, le domaine temporel doit être supérieur ou égal à la période sur laquelle porte l'étude, majorée de la durée maximale entre la période de début et chaque événement[3].

Ce domaine temporel doit être déterminé par le demandeur et peut être calculé sur la base d'input qui peut être obtenu auprès des sources authentiques. Les sources authentiques devront prévoir une marge suffisante afin que cet input ne soit pas une date exacte.

[1] P.ex. la fonction SecureRandom d'un ordinateur permettant d'obtenir une répartition aussi uniforme que possible.
[2] mod : calcul modulo fournit le reste de la division d'un nombre entier par un autre nombre entier
[3] P.ex. une étude pour laquelle la date d'incidence est prise comme critère de sélection et la période sur laquelle porte l'étude s'élève à 3 ans. Lorsque la série de données comprend des dates d'évènements qui peuvent se situer pour un intéressé jusqu'à 1 an après la date d'incidence, le domaine temporel sera donc (3 + 1) an, c'est-à-dire 1461 jours.

Dans l'hypothèse où il n'est pas possible d'obtenir un input de la source authentique pour le calcul du domaine temporel, le domaine temporel peut prendre cours à la date de début de la période sur laquelle porte l'étude et se terminer le jour de l'extraction des données par les sources authentiques.

Risques

Les risques suivants sont pris en compte et des méthodes de protection sont décrites :

1. La détermination de dates exactes sur la base des dates de référence :
    a. La série de données finale ne peut pas contenir d'information sur la date de référence, de sorte que ce calcul ne soit pas possible. Le destinataire ne peut pas non plus obtenir ces dates de référence d'une autre façon.
    b. La présente proposition évite la transmission de dates exactes pour le calcul de durées. Etant donné qu'il s'agit de calculs relatifs, la date de référence n'est pas nécessaire pour obtenir un résultat correct.
2. L'estimation de dates réelles sur la base de NewDates :
    a. Grâce à l'utilisation du calcul modulo, chaque date réelle est projetée vers un point aléatoire du domaine temporel qui a été défini dans la demande. La position de NewDate est la somme de la date de référence et de la date réelle. Grâce à l'application du calcul modulo, il n'est pas possible de déterminer si la somme de ces deux dates est inférieure ou supérieure au nombre maximal de jours dans le domaine déterminé.
3. La perte de qualité des informations en raison du décalage des dates :
    a. Les calculs prévus sur la base de NewDates concernent des durées. Etant donné qu'il s'agit d'un calcul relatif et que le décalage est identique pour les deux dates, il n'y a pas de perte de qualité pour ce calcul. L'établissement de recherche devra cependant adapter ses algorithmes à cette proposition.
4. La perte d'intégrité des informations obtenues suite à l'application du calcul modulo :
    a. Lorsqu'une durée entre p.ex. l'incidence et un événement associé est supérieure de manière absolue au domaine temporel, la durée calculée risque de différer un multiple de la durée du domaine temporel. Ceci doit être vérifié lors de la détermination du domaine temporel.

Exemple

A titre d'illustration, voici un exemple élaboré à l'aide des paramètres aléatoires suivants :

Date de début de la période de la recherche :	01/01/2010 (00:00)
Date de fin de la période de la recherche :	31/12/2019 (23:59)
Durée maximale entre le premier et le dernier évènement pour chaque intéressé :	2 ans
Date réelle 1	15/02/2016
Date réelle 2	13/07/2018
Date réelle 3	20/10/2020

Calcul du domaine temporel

Le domaine temporel est la somme de la durée de la période de la recherche et de la durée maximale entre un premier évènement et le dernier évènement pour chaque intéressé.

En l'occurrence :

Durée de la période de la recherche :	3652 jours (10 ans)
Durée maximale entre le premier et le dernier évènement pour chaque intéressé :	732 jours[4] (2 ans)
Durée du domaine temporel (MaxDays) :	4384 jours

Sachant que le domaine temporel :

débute le :	01-01-10 (00:00)
se termine le :	02-01-22 (00:00)

Calcul de durée

Les durées, calculées à partir de dates réelles, sont :

Durée Date réelle 1 - Date réelle 2 :	879 jours
Durée Date réelle 1 - Date réelle 3 :	1709 jours

Pour déterminer les NewDates, il convient de choisir d'abord une date de référence. Ceci est réalisé sur la base d'un nombre naturel aléatoire inférieur à MaxDays. Dans cet exemple, la RefDate **956** (14/8/2012) est choisie.

NewDate = (date de début domaine temporel) + ((Date réelle domaine temporelle) + (Ref date)) mod (MaxDays). Ce qui donne les résultats suivants :

NewDate1	(01/01/2010) + (((15/02/2016) – (01/01/2010)) + 956)mod(4384)) jours= (01/01/2010) + ((2236 + 956)mod(4384)) = (01/01/2010) + 3192 = **28/9/2018**
NewDate2	(01/01/2010) + (((13/07/2018) – (01/01/2010)) + 956)mod(4384)) jours= (01/01/2010) + ((3115 + 956)mod(4384)) = (01/01/2010) + 4071 = **23/02/2021**
NewDate3	(01/01/2010) + (((20/10/2020) – (01/01/2010)) + 956)mod(4384)) jours= (01/01/2010) + ((3945 + 956)mod(4384)) = (01/01/2010) + ((4901)mod(4384)) =

[4] Pour simplifier le calcul, il a été opté ici pour 366 jours par an. De cette manière, l'effet d'une année bissextile est toujours pris en compte.

(01/01/2010) + 517 = **02/06/2011**

Pour le calcul des durées effectives sur la base de NewDates, la formule suivante est appliquée :

Duration = (NewDate2 – NewDate1 + MaxDays) mod (MaxDays)

Ce qui donne pour cet exemple le résultat suivant :

Durée NewDate1 – NewDate2 :

((23/02/2021 – 28/09/2018)+ 4384)mod(4384) =
(879 + 4384)mod(4384) =
(5363)mod (4384) =
879 jours

Durée NewDate1 – NewDate3 :

((02/06/2011 – 28/09/2018)+ 4384)mod(4384) =
(-2675 + 4384)mod(4384) =
(1709)mod (4384) =
1709 jours

De manière graphique, on obtient ce qui suit :