Illustrating Evidence Based Anonymization

Nobuaki HOSHINO[*]

January 2016

Abstract

To publish a data set we must ensure that included individuals are unidentifiable. Nevertheless this unidentifiability is often judged rather subjectively. The author has proposed objectifying this practical judge by the statistical estimation of a model, where an unknown parameter describes uncertainty on identification. This idea exploits observation as a statistical evidence. To promote this evidence based method, the present article demonstrates its applicability to the decision of the acceptable range of disclosure risk on Japanese Anonymized Data.

Key words: Population unique, Privacy, Statistical Disclosure Control.

1 Introduction

Confidentiality or privacy reasons let laws allow no individual to be identified when data are published. Many techniques to prevent such identification are known as anonymization; see e.g. Willenborg and de Waal (1996, 2000). After anonymizing a data set, data publishers need to decide whether the set is identifiable or not. This decision, however, has been made in a rather subjective manner.

Technical efforts to formalize this decision result in many measures of (re-)identification risk, but no objective method to decide a threshold of those measures seems known. This is so because those researches authorize this decision to depend on a personal preference under a tradeoff between re-identification risk and utility of data; see e.g. Duncan et al. (2001).

However, identifiability does not depend on a personal preference: When data are identifiable they are so regardless of the preference of their publisher. Also laws say nothing about the utility of data. In other words, no identifiable data set is publishable even if they are highly useful. Therefore in practice the true goal of anonymization is not to take the balance of risk and utility but to maximize utility within an acceptable range of risk. This goal has been claimed by Dalenius (1977) and others, and for a practitioner the decision of the threshold of identification risk is the primary issue.

This important decision should be objective and explicable. To achieve this clarity the present article employs Hoshino's (2013) statistical model of identification with an unknown threshold of risk, and estimates it by observing whether published data have been identified

[*] School of Economics, Kanazawa University, Kakuma-machi, Kanazawa 920-1192, Japan. E-mail: hoshino@kenroku.kanazawa-u.ac.jp

or not. These observations are statistical evidence that carries information on the threshold of identification risk. Thus we call our approach “Evidence Based Anonymization (EBA)”.

In this way an estimated threshold depends on the measured elements of risk. Our risk measure only accounts for the difference of data and neglect institutional effects as typical risk measures do. Institutional effects, however, seem to exist since the degree of anonymization empirically depends on the qualifications of users of data: A public use file tends to be more anonymized than a scientific use file. Therefore in order to control institutional effects that may exist, we restrict our sample space to the cases of each institution. In other words, the threshold of our risk measure can differ among institutions. Such a difference arises from the difference of pooled institutional effects.

Restricting our sample space is advantageous in that we do not have to decompose institutional effects. This decomposition is virtually impossible due to very limited information on the latent ability of identification that a society possesses.

To exemplify EBA, the present article estimates the risk threshold of Japanese Anonymized Data of surveys conducted by Ministry of Internal Affairs and Communications (MIAC). Anonymized Data are defined by Statistics Act so that no individual shall be identified. Global recoding and record suppression are mainly used to satisfy this definition.

MIAC’s Anonymized Data have been available on 4 surveys (Employment Status Survey, Housing and Land Survey, National Survey of Family Income and Expenditure, and Survey on Time Use and Leisure Activities) from 2009, on Labour Force Survey from 2011, and on Population Census from 2013. The Japanese society has not recognized any identification of an individual contained in these data as of 2015.

Anonymized Data are provided for academic research and advanced education under a license. The number of past users of these data varies among surveys, but in total about 30 applicants have passed the same review procedure for use in each year as of 2013. We consider that these cases of using Anonymized Data share the same threshold of identification risk, since they are under the same institutional measures against identification.

The present article is organized as follows. Section 2 explains our statistical model that links our risk measure to the observation of identification. Section 3 explains the detail of evaluating our risk measure. Section 4 estimates the threshold of identification risk of Anonymized Data. Section 5 concludes.

2 Statistical model of identification

This section explains Hoshino’s (2013) method to decide whether a measured identification risk is acceptable or not. The first subsection technically describes unidentifiability. The second subsection presents a model whose parameter expresses uncertainty on identification. The third subsection statistically estimates this paramter, by which the acceptable range of our risk measure is determined.

2.1 Definition of identifiability

An effort on modeling identification can be seen in Marsh et al. (1991). They argue that the probability of identification is the product of the following probabilities:

$$\Pr(\text{actual identification}) = \Pr(\text{success of identification}|\text{trial of identification}) \Pr(\text{trial of identification}). \tag{1}$$

In eq. (1), the event of "actual identification" is regarded as the joint event of "success of identification" and "trial of identification". This discrimination between "actual identification" and "success of identification" corresponds to different legal concepts of anonymity or unidentifiability.

Absolute anonymity, using a German legal term, is a state where the possibility of identification is eliminated with no doubt. We regard this state as equivalent to a state that

$$\Pr(\text{success of identification}|\text{trial of identification}) = 0. \tag{2}$$

De facto anonymity, which is also a German legal term, is a state where the cost of identification dominates the benefit of identification. In this case the probability of the trial of identification should be low, and thus we regard this state as equivalent to a state that Pr(actual identification) is low.

The present article focuses upon the assessment of the absolute anonymity, because Japan Law seems to define confidentiality as such. Consequently we evaluate whether the conditional probability of "success of identification" given "trial of identification" is zero or not.

Marsh et al. (1991) regard the success of identification as the result of matching betweeen a published file and an outer file owned by an attacker (who tries to identify an individual). They accordingly propose the following factorization:

$$\Pr(\text{success of identification}|\text{trial of identification}) = \Pr(a) \Pr(b|a) \Pr(c|a, b) \Pr(d|a, b, c), \tag{3}$$

where the events from a to d are

(a) On the attribute of the same individual, both a published file and an outer file record the same value (i.e. no misclassification etc.).

(b) A published file contains an individual.

(c) An individual is a population unique.

(d) A population unique is verified to be so.

If we can evaluate the probabilities of the right hand side of eq. (3), we can obtain the conditional probability of "success of identification" given "trial of identification". However, the evaluation of these probabilities by Marsh et al. (1991) is not convincing from a modern point of view.

As discussed in Section 3, $\Pr(a, b, c)$ depends on information that an attacker currently knows, yet $\Pr(d|a, b, c)$ should depend on information that an attacker does not currently know. An attacker may be able to collect additional information to verify a population unique. Such currently nonexistent information is unobservable and hard to estimate. Hence the author considers that no one can plausibly evaluate $\Pr(d|a, b, c)$.

Now let us be reminded that we just would like to know whether eq. (2) holds or not. This evaluation is far easier than to evaluate the conditional probability of "success of identification" given "trial of identification".

Therefore we rewrite eq. (3) as

$$\Pr(\text{success of identification}|\text{trial of identification}) = \Pr(a, b, c)\Pr(d|a, b, c). \tag{4}$$

Then we can see that eq. (2) holds if and only if at least one of $\Pr(a, b, c)$ and $\Pr(d|a, b, c)$ is zero. On data for scientific purposes, $\Pr(a, b, c)$ is usually positive. Consequently our usual assessment on unidentifiability reduces to a decision whether $\Pr(d|a, b, c)$ equals zero or not. Since the direct evaluation of $\Pr(d|a, b, c)$ is hopeless, we will estimate whether $\Pr(d|a, b, c)$ equals zero or not.

2.2 Model for discerning identifiability

From our argument so far, we would like to discern whether

$$\Pr(d|a, b, c) = 0 \tag{5}$$

or not, since eq. (5) is sufficient for an unidentifiable state.

To this goal we note that $\Pr(d|a, b, c)$ is subject to the event of (a, b, c), and $\Pr(a, b, c)$ can be evaluated since it only depends on existent information. The increment of $\Pr(a, b, c)$ implies that more information about population uniques is published. The more information exists, more easier the verification of a population unique should become. Hence the conditional probability of d given (a, b, c) should be monotonically increasing as $\Pr(a, b, c)$ increases. If so, there exists nonnegative β such that

$$\Pr(a, b, c) \leq \beta \Leftrightarrow \Pr(d|a, b, c) = 0. \tag{6}$$

Then we conclude that the assessment of identifiability reduces to the evaluation of $\Pr(a, b, c)$, since eq. (2) is tantamount to $\Pr(a, b, c) = 0$ or $\Pr(d|a, b, c) = 0$.

In the model (6), $\Pr(a, b, c)$ can be interpreted as the easiness of identification. This is a type of re-identification risk measure, and its threhold β is unknown. We decide it by statistical estimation in the following.

2.3 Observational model of identification

For the statistical estimation of β in eq. (6), we need an observation that carries information on β. Hence we would like to observe the event of d or the success of identification. However, a society may not always recognize such an event; a successful attacker may hide. Therefore we discriminate "actual identification" from its social recognition.

Let a random variable X be 1 when "actual identification" is socially recognized, and 0 otherwise. That is,

$$\Pr(X = 1) = \Pr(\text{recognized}|\text{actual identification})\Pr(\text{actual identification}).$$

Then from eq. (1) and eq. (3),

$$\begin{aligned}\Pr(X = 1) &= \Pr(\text{recognized}|\text{actual identification}) \\ &\quad \times \Pr(a, b, c, d)\Pr(\text{trial of identification}). \end{aligned} \tag{7}$$

Further, let us write the evaluated value of $\Pr(a, b, c)$ as γ, and write

$$p(\gamma) = \gamma \Pr(d|a, b, c)\Pr(\text{recognized}|\text{actual identification})\Pr(\text{trial of identification}). \tag{8}$$

Then

$$\Pr(X=1) = \begin{cases} p(\gamma) & \text{if } \gamma > \beta \\ 0 & \text{if } \gamma \leq \beta. \end{cases} \tag{9}$$

If $p(\gamma)$ is positive, the observed value of X carries information on β, and we can estimate β from X's. Actually $p(\gamma)$ is positive when both

$$\Pr(\text{recognized}|\text{actual identification})) > 0 \tag{10}$$

and

$$\Pr(\text{trial of identification}) > 0 \tag{11}$$

hold. The first condition (10) should be satisfied because an attacker has an incentive to show off their success of identification. Also hiding through is not always possible. The second condition (11) should also be satisfied because of a potential incentive to do so. Hence we regard that $p(\gamma)$ is positive. It is worth mentioning that we assume no specific form of $p(\cdot)$.

Suppose that there are n past experiences of publishing anonymized data. We regard these as independent samples from the model (9). For the i-th, $i = 1, 2, \ldots, n$, sample we measure $\Pr(a, b, c) = \gamma_i$ and observe the social recognition of actual identification $X_i = x_i$. Write the likelihood of the observations as $\ell(\beta)$. To simplify our argument we assume that $\gamma_1 > \gamma_2 > \cdots > \gamma_n$.

Now we consider the maximum likelihood estimator $\hat{\beta}$ of the threshold. If there exists an integer m such that $x_{m-1} = 1, x_m = x_{m+1} = \cdots = x_n = 0$, then $\ell(\beta) = 0$ for $\beta \geq \gamma_{m-1}$, $\ell(\beta) \propto p(\gamma_{m-1})$ for $\gamma_{m-1} > \beta \geq \gamma_m$, and $\ell(\beta) \propto p(\gamma_{m-1}) \prod_{j=m}^{i}(1 - p(\gamma_j))$ for $\gamma_i > \beta \geq \gamma_{i+1}, i \geq m$. Hence $\gamma_{m-1} > \hat{\beta} \geq \gamma_m$ because $p(\gamma)$ is positive. If there exists no social recognition of actual identification, then $\hat{\beta} \geq \gamma_1$.

In general we denote the lowest easiness of identification among samples with social recognition of actual identification by γ^-. If there has been no such recognition, let γ^- be 1. Also among samples with the easiness that is lower than γ^- we denote the highest easiness of identification by γ^+. Then $\gamma^+ \leq \hat{\beta} < \gamma^-$.

3 Measuring disclosure risk

To substantiate our theoretical model of the previous section, we have to establish the method of evaluating $\Pr(a, b, c)$, which is explained in this section following Hoshino (2013). The first subsection clarifies the policy of selecting key variables or quasi-identifiers. Under this policy the second subsection decomposes $\Pr(a, b, c)$ to the product of $\Pr(a), \Pr(b|a)$ and $\Pr(c|a, b)$, and explains the method of evaluating each probability.

3.1 How to select key variables

First we discuss the policy of selecting key variables for matching, on which the number of population uniques heavily depends. Because EBA compares $\Pr(a, b, c)$ among cases, the policy must be fixed.

Few arguments exist on the formal selection of key variables. For example, Elliot et al. (2010, 2011) claim a comprehensive survey of existent information about individuals in a society for this purppose. The author never denies the importancd of such information, but their argument does not directly result in the best selection of key variables. Fung et al. (2010) describe the selection of key variables as "an open problem".

The present article selects key variables to best estimate β. Existing researches can not optimize the selection because they do not consider the aftermath of evaluating population uniques.

Suppose that there are k variables in a published file. Then there are 2^k ways to select key variables in theory. The number of population uniques can be evaluated in each way, and we write the order statistics of these numbers as $u_{(1)} \leq u_{(2)} \leq \cdots \leq u_{(2^k)}$. Then the issue of the selection of key variables is nothing but the selection of a rank among $(1, 2, \ldots, 2^k)$, over which attackers are distributed subject to their knowledge about individuals.

For given data we evaluate $\Pr(a, b, c)$ at a selected rank r, and compare it with γ^+, which is also evaluated under the same policy of selecting key variables. Suppose that $\Pr(a, b, c)$ at the r-th rank is smaller than γ^+. Then, for fixed $\Pr(a, b)$, the given data should be safe against attackers who lie on ranks smaller than r, since $u_{(i)} \leq u_{(r)}$ for $i \leq r$. The given data, nevertheless, have no evidence of safety against attackers who lie on ranks larger than r.

Thus one might think that we should select the largest rank: 2^k. However, an attacker may not exist on the 2^k-th rank. If so, an observed X of eq. (9) carries no information on the safety of $\Pr(a, b, c)$ at the 2^k-th. Hence, considering the distribution of attackers over the ranks, we should select the largest rank on which an attacker exists.

The best way to select key variables has been described theoretically, but in practice, the distribution of attackers is unknown. Therefore we have to estimate the maximum of the distribution of attackers over ranks. The precise estimation of a maximum is, however, known to be difficullt, the theory of extreme values might be usable though. Also an errorneous estimate of the maximum rank leads to unstable $\hat{\beta}$. Hence, as a second best way, we should estimate a percentile, which is less difficult. For example it should be more practical to estimate the 99th percentile of the distribution of attackers, as in the case of financial risk called Value at Risk (VaR).

Unfortunately the quantitative evaluation of such a percentile is virtually impossible since we can not observe the distribution of attackers. Hence we select key variables whose information is publicly known. This policy is common in practice, which actually implies that some large percentile is estimated.

Our policy sacrifices anonymization's grip on the strongest attackers, but other institutional protection may suffice. For example, since the strongest attackers should be conspicuous, a data publisher may be able to reject their request for a scientific use file. It censors the right tail of the distribution of attackers. A penal code should be effective even in the case of a public use file.

3.2 Risk measured to control most

This section describes the method of measuring $\Pr(a), \Pr(b|a)$ and $\Pr(c|a, b)$ under our policy of measuring risk at a large percentile.

3.2.1 Measuring $\Pr(a)$

The attribute of an individual may be differently recorded between a published file and an outer file. Marsh et al. (1991) ascribe this difference to an error in recording or a change of an attribute with the passage of time. A perturbation technique such as swapping can also cause this difference.

If at least one key variable of an individual is affected by these causes, then the event of a does not occur. Therefore Marsh and others claim that the increment of key variables tends to decrease $\Pr(a)$; Shlomo and Skinner (2010) give a numerical example of this kind.

Nevertheless they neglect the possibility of the correction of such differences. A record-linkage-like technique can correct them especially when a unique individual lies in a sparse space. Because this sparsity emerges when key variables increase, the increment of key variables does not necessarily decrease $\Pr(a)$. Hence we do not relate $\Pr(a)$ to the number of key variables.

Consequently we evaluate $\Pr(a) = 1$ for an unperturbed file. This evaluation does not imply no error in recording. The rate of errors, which is uncontrolable by a data pulisher, is a part of uncertainty on identification. Hence we consider that the rate of errors too is described by β.

The effect of perturbation should be evaluated casewise. The present article does not deal with a perturbed file, and thus we do not argue further.

3.2.2 Measuring $\Pr(b|a)$

Following Marsh et al. (1991) we evaluate $\Pr(b|a)$ as the ratio of the size of a published file to the correspoinding population size.

3.2.3 Measuring $\Pr(c|a,b)$

Marsh et al. (1991) defines $\Pr(c|a,b)$ by a ratio of the number of population uniques to its population size. The evaluation of this ratio usually involves estimating the number of population uniques, which is not straightforward.

On this estimation we employ Hoshino's (2001) method that exploits Pitman's (1995) sampling formula. Our method is advantageous in that it does not require tailored modeling for each data set. It is thus suitable for comparing many data sets with the same standard.

Regression is a common way to estimate the number of population uniques, but as Skinner and Shlomo (2008) address, such an inference may be sensitive to the specification of a model. Therefore regression is unsuitable for our comparison of estimates.

Another advantage of our method is its applicability to sparse data. Many key variables are likely to be selected under our policy. Then individuals are distributed over the high dimensional space of these key variables, and a frequency on a location in this space tends to be zero, which implies sparse data. If most of these frequencies are zero, such a location indexed by the values of key variables has little information on the number of uniques. That is, key variables are useless to estimate the number of uniques: Regression fails. Our method still works.

4 Risk of Anonymized Data

This section demonstrates the estimation of β for the cases of Anonymized Data. Results are to be presented only orally.

5 Concluding remarks

The implication of our argument in Section 2 is clear: A given data set is publishable if its $\Pr(a, b, c)$ does not exceed γ^+, since it has a statistical evidence of unidentifiability. This γ^+ is estimated in Section 3 for Anonymized Data provided by MIAC.

By these arguments the present article illustrates one method to objectively decide whether given data are identifiable or not. Subjective decision may employ past experiences implicitly, yet our method explicitly employs past experiences as a statistical evidence.

What if there has been no past example that can be an evidence? Then begin with publishing apparently safe data; a clinical trial decides the threshold of some dose by gradually increasing risk. The idea of evidence based decision originates in medicine. It can also be applied to anonymization.

Acknowledgements

This research has been supported by Kakenhi grant from the Japanese Ministry of Education, Culture, Sports, Science and Technology.

Anonymized Data used in the present article have been provided by the National Statistics Center of Japan, subject to Statistics Act.

References

[1] Dalenius, T. (1977). Towards a methodology for statistical disclosure control. *Statistisk Tidskrift*, **15**, 428–444.

[2] Duncan, G., Keller-McNulty, S.A. and Stokes, S.L. (2001) Disclosure Risk vs. Data Utility: The R-U Confidentiality Map. Technical Report 121, National Institute of Statistical Sciences, Durham, North Carolina.

[3] Elliot, M., Lomax, S., Mackey, E. and Purdam, K. (2010) Data Environment Analysis and the Key Variable Mapping System. *Privacy in Statistical Databases*, Domingo-Ferrer, J. and Magkos, E. (Eds.), LNCS 6344, 138–147, Springer-Verlag, Berlin Heidelberg.

[4] Elliot, M., Mackey, E. and Purdam, K. (2011) Formalizing the Selection of Key Variables in Disclosure Risk. *Int. Statistical Inst.: Proceedings of the 58th World Statistical Congress*, 2777–2784.

[5] Fung, B.C.M., Wang, K., Fu, A.W.C and Yu, P.S. (2010) *Introduction to Privacy-Preserving Data Publishing*, CRC Press, New York.

[6] Hoshino, N. (2001) Applying Pitman's Sampling Formula to Microdata Disclosure Risk Assessment, *Journal of Official Statistics*, **17**, 499–520.

[7] Hoshino, N. (2013) *Evidence Based Anonymization.* Discussion Paper No.21, Faculty of Economics and Management, Kanazawa University. (In Japanese.)

[8] Marsh, C., Skinner, C., Arber, S., Penhale, P., Openshaw, S., Hobcraft, J., Lievesley, D. and Walford, N. (1991) The Case for a Sample of Anonymized Records from the 1991 Census. *Journal of the Royal Statistical Society*, Series A, **154**, 305–340.

[9] Pitman, J. (1995) Exchangeable and Partially Exchangeable Random Partitions. *Probability Theory and Related Fields*, **102**, 145–158.

[10] Shlomo, N. and Skinner, C. (2010) Assessing the protection provided by misclassification-based disclosure limitation methods for survey microdata. *Annals of Applied Statistics*, **4**, 1291–1310.

[11] Skinner, C. and Shlomo, N. (2008) Assessing Identification Risk in Survey Microdata Using Log-Linear Models. *Journal of the American Statistical Association*, **103**, 989–1001.

[12] Willenborg, L. and de Waal, T. (1996). *Statistical Disclosure Control in Practice* , Lecture Notes in Statistics 111, Springer, New York.

[13] Willenborg, L. and de Waal, T. (2000). *Elements of Statistical Disclosure Control.* Lecture Notes in Statistics 155, Springer, New York.

季節調整プログラム X-13ARIMA-SEATS について

高岡　慎

琉球大学法文学部

2016 年 1 月 29 日

科研費プロジェクト『経済統計・政府統計の理論と応用からの提言』カンファレンス

はじめに

■ 本報告の内容

1. X-11 から X-13ARIMA-SEATS へ
2. RegARIMA モデルの概要
3. TRAMO-SEATS による季節調整
4. X-13ARIMA-SEATS
5. 調整結果の比較
6. まとめ

1. X-11からX-13ARIMA-SEATSへ

■ X-11

- Shiskin, Young, and Musgrave(1967)
- 移動平均フィルタの連続的な適用による時系列の分解
- 単純な対称移動平均フィルタ
- ヘンダーソン移動平均フィルタ
- 端点付近では非対称フィルタ（マスグレーブ法）

1. X-11からX-13ARIMA-SEATSへ

■ X-11-ARIMA

- カナダセンサス局（Dagum(1988)）
- RegARIMAモデルの導入
  ⇒RegARIMAモデルによる予測値で時系列を延長
- 対称移動平均フィルタの適用

1. X-11 から X-13ARIMA-SEATS へ

■ X-12-ARIMA

- Findley, Monsell, Bell, Otto, Chen(1998)
- RegARIMA モデルの使用
  ⇒ 異常値・レベルシフトなどの変動を回帰変数として処理
- モデルによる系列の延長
- X-11 フィルタによる処理
- 事後診断機能など、安定した季節調整のための様々な改良

1. X-11 から X-13ARIMA-SEATS へ

■ TRAMO-SEATS

- スペイン銀行（Maravall(1995) その他）
- RegARIMA モデルの利用
  ⇒ 異常値・レベルシフトの処理と時系列構造の特定
- TRAMO パートでのモデル選択
  ⇒ 単位根検定と情報量規準による選択
- SEATS パートでの信号抽出に基づく季節調整
  ⇒WK(Wiener-Kolmogorov) フィルタによる時系列の分解

1. X-11からX-13ARIMA-SEATSへ

■ X-13ARIMA-SEATS

- X-12-ARIMAとTRAMO-SEATSの統合
- TRAMOパートをRegARIMAモデルの自動モデル選択コマンドとして内蔵
- SEATSの機能を全て実装
- 季節調整処理ではX-11法とSEATSを選択可能

2. RegARIMAモデルの概要

■ RegARIMAモデル

原系列y_tが

$$y_t = \sum_i \beta_i x_{it} + z_t$$

のような線形回帰モデルの形式で表され、z_tが季節ARIMAモデルに従うとき、y_tのモデルをRegARIMAモデルとよぶ。一般的な表記は

$$\phi(B)\Phi(B^s)(1-B)^d(1-B^s)^D\left(y_t - \sum_i \beta_i x_{it}\right) = \theta(B)\Theta(B^s)a_t$$

となる。Bはバックシフトオペレータ、$\phi(B), \Phi(B^s), \theta(B), \Theta(B^s)$は$B$の多項式、$a_t$はホワイトノイズ。

2. RegARIMA モデルの概要

■ モデルの選択

- **pickmdl** コマンドによりモデルの次数を自動選択することが可能。
- **pickmdl** コマンドは X-12-ARIMA では **automdl** という名称だったが、X-12-ARIMA の最終バージョンでは TRAMO のモデル選択法 (後述) が導入され、従来の選択アルゴリズムは **pickmdl** に名称が変更された。
- 候補となるモデルのインサンプルでの backcast error と forecast error が計算され、一定の規準を満たしているモデルが選択される。
- 階差次数も同時に選択するために、予測誤差に基づく経験的な統計量が採用されている。
- 候補が全てリジェクトされる場合もあり、使いにくい点があった。

2. RegARIMA モデルの概要

■ モデルの推定

繰り返し一般化最小二乗法 (IGLS) による推定

(1) 原系列と説明変数系列の両方に必要な階差を適用

(2) 与えられた AR と MA 母数に対して回帰係数を一般化最小二乗法 (GLS) で推定

(3) 回帰モデルの母数 β_i の値を所与として最尤法により ARMA モデルの係数を推定

(4) (2) と (3) のプロセスを収束するまで反復

3. TRAMO-SEATSによる季節調整

■ TRAMO-SEATSの概要1

- スペイン銀行により開発されたモデルベースの季節調整法 (Gomez and Maravall(1996))
- TRAMO(Time series Regression with ARIMA noise, Missing Observation and Outliers) パートと SEATS(Signal Extraction in ARIMA Time Series) パートに分かれており、前者では原系列に適用すべき RegARIMA モデルの選択・推定、後者では信号抽出による季節調整に関する処理が行われる。

3. TRAMO-SEATSによる季節調整

■ TRAMO-SEATSの概要2

- RegARIMA モデルは

  TRAMO パートの事前調整
  - ・RegARIMA モデルの適用
  - ・外れ値の処理
  - ・欠損値の処理
  - ・前方予測と後方予測の追加

  などの処理に利用される。
- 事前調整が行われた後で、WK フィルタによる時系列の分解が行われる。
- 全体の処理の流れは X-12-ARIMA のプロセス類似しているが、利用するフィルタが異なる。

3. TRAMO-SEATSによる季節調整

■ 信号抽出による時系列の分解1

系列$\{X_t\}$がシグナル(S_t)とノイズ(N_t)の和

$$X_t = S_t + N_t$$

であるとする。シグナルとノイズはいずれも定常で、かつ互いに独立。

$\{X_t\}$のみが観察可能であるとき、X_tを利用してS_tを推定する。

線型な推定量

$$\hat{S}_t = \sum_{j=-\infty}^{\infty} \varphi_j X_{t-j} = \varphi(B) X_t$$

を考え、最適なウェイトを考える。

3. TRAMO-SEATSによる季節調整

■ 信号抽出による時系列の分解2

平均2乗誤差

$$E\left[(S_t - \hat{S}_t)^2\right]$$

を最小にするものを最適なウェイトと考えると、

$$\hat{S}_t \quad = \quad \varphi(B) X_t = \frac{\sigma_s^2}{\sigma^2} \frac{\psi_s(B)\psi_s(B^{-1})}{\psi(B)\psi(B^{-1})} X_t$$

が得られる。

⇒WK(Wiener-Kolmogorov)フィルタ

3. TRAMO-SEATSによる季節調整

■ 信号抽出による時系列の分解3

ただし

$$
\begin{aligned}
X_t &= \psi(B)\epsilon_t, \quad \epsilon_t \sim WN(0,\sigma^2) \\
S_t &= \psi_s(B)\epsilon_{s,t}, \quad \epsilon_{s,t} \sim WN(0,\sigma_s^2) \\
N_t &= \psi_n(B)\epsilon_{n,t}, \quad \epsilon_{n,t} \sim WN(0,\sigma_n^2)
\end{aligned}
$$

としている。

X_t の時系列モデルを特徴づける $\psi(B)$ をモデル選択プロセスから特定し、幾つかの制約条件の下で $\psi(B)$ から $\psi_s(B)$ を導けば、

$$
\hat{S}_t = \frac{\sigma_s^2}{\sigma^2}\frac{\psi_s(B)\psi_s(B^{-1})}{\psi(B)\psi(B^{-1})}X_t
$$

の右辺のウェイトが導かれる。

4. X-13ARIMA-SEATS

■ X-12とTRAMO-SEATSの統合

- 後述するX-13ARIMA-SEATSのAUTOMDLコマンドは、ほぼTRAMOパートでのモデル選択の処理を踏襲している。
- 信号抽出による季節調整を実行するSEATSパートはX-13ARIMA-SEATSでは**seats**コマンドとして実装されている。
- RegARIMAモデルによる処理は共通化され、回帰変数の処理や系列の延長に利用される。
- X-13ARIMA-SEATSでは季節調整の方法として、X-11フィルタとWKフィルタのいずれかを選択可能。
- **seats**コマンドによりWKフィルタを選択した場合は、RegARIMAモデルの推定結果がフィルタの導出にも利用される。

4. X-13ARIMA-SEATS

■ 新たに追加されたコマンド

- AUTOMDL
  ⇒TRAMO と同様の自動モデル選択の実行
- PICKMDL
  ⇒X-12-ARIMA の自動モデル選択プロセス
- SEATS
  ⇒SEATS の季節調整法の実行
- SPECTRUM
  ⇒ 季節性や曜日効果の事後診断を行うためにスペクトラムを出力
- FORCE
  ⇒ 原系列と季節調整系列で年間の集計値が一致するように制約をかけるオプションコマンド

4. X-13ARIMA-SEATS

■ AUTOMDL コマンド

(1) **デフォルトモデルの推定**
⇒「エアラインモデル」による回帰変数について予備的な推定

(2) **階差および季節階差の次数の特定**
⇒(1) の残差系列に対する単位根のチェック

(3) **ARMA 部分の次数の特定**
⇒ 特定された階差を適用し、ARMA 部分の次数を BIC により特定

(4) **特定されたモデルとデフォルトモデルの比較**

(5) **最終チェック**

4. X-13ARIMA-SEATS

■ (1) デフォルトモデルの推定

- ARIAM部分を「エアラインモデル」(0 1 1)(0 1 1)に固定し、回帰部分の係数を推定する。
- 各回帰係数についてt検定で有意性を確認する。
  ⇒ サンプル数に応じたCV(critical value)を用いる。
- **outlier** コマンドで外れ値の自動選択を指定している場合はここで実行。
- 外れ値ダミーを加えても曜日効果、イースター効果、定数項が有意かチェック。
- 回帰変数が特定されると、モデルの残差に対するLjung-BoxのQ統計量が計算される。（後の手順で使用）
- 回帰変数の効果を除いた系列 **linearized series** が計算される。

4. X-13ARIMA-SEATS

■ (2) 階差および季節階差の次数の特定

- **linearized series** に対して適用すべき階差を特定する。
- 一般的な単位根検定ではなく、Hannan-Rissanen法などによって推定されたARMAモデルのAR係数を規定の値と比較することで単位根の有無を判定。
- 単位根がある場合はそれに相当する階差を適用し、再度ARMAモデルを推定する。
- 単位根が検出されなくなるまで手順を繰り返す。

4. X-13ARIMA-SEATS

■ (3)ARMA 部分の次数の特定

- **linearized series** に対して特定された階差操作を適用した系列について、ARMA 次数を特定する。
- 次数の上限を設定し、BIC の比較により選択する。
- 選択 S れたモデルがデフォルトモデルと異なる場合は、デフォルトモデルとの比較を行う。（手順 (4)）

4. X-13ARIMA-SEATS

■ (4) 特定されたモデルとデフォルトモデルの比較

- P_A と P_D をそれぞれ自動選択モデルとデフォルトモデルの Q 統計量の p 値とし、

  $$Q_A = 1 - P_A, \quad Q_D = 1 - P_D$$

  とする。
- Q_A と Q_D の値や相対的な大小関係などについての一定の規準により、自動選択モデルとデフォルトモデルのうちどちらかが選ばれる。
- 選ばれたモデルの Q が 0.975(デフォルト) を超えている場合にはモデルが不適切と判断 ⇒ 外れ値の境界 CV を少し小さくし、(1) の外れ値の検出からやり直す。

4. X-13ARIMA-SEATS

■ (5) 最終チェック

- 手順 (4) をクリアしたモデルの最終チェックが行われる。
- AR 部分の単位根の確認

  ⇒ 特性根の絶対値が 1.05 以下なら単位根と判断し、AR 次数を減らし階差を増加させる。
- MA 部分の単位根の確認

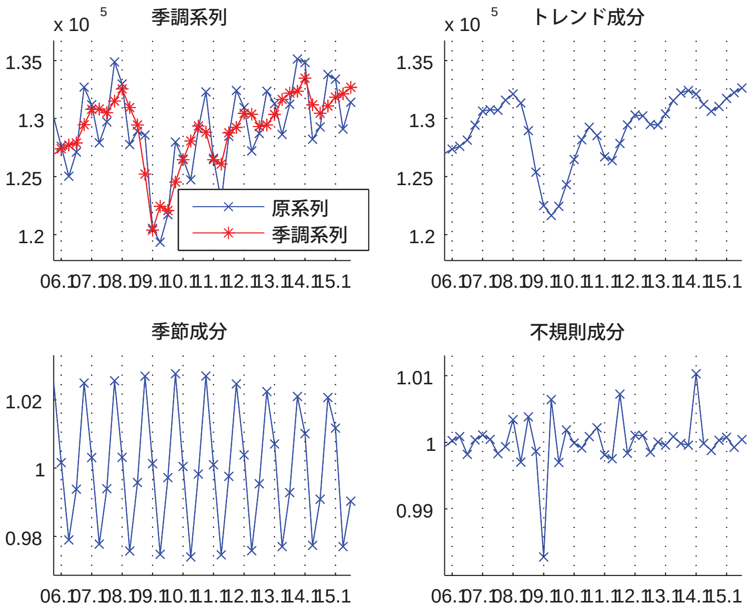
  ⇒ モデルの反転可能性を確認
- ARMA パラメータの有意性

  ⇒AR、季節 AR、MA、季節 MA のそれぞれの最大次数のパラメータが有意かどうかをチェック

  ⇒ 有意でないものがあれば、CV を小さくし手順 (1) の外れ値の検出からやり直す。

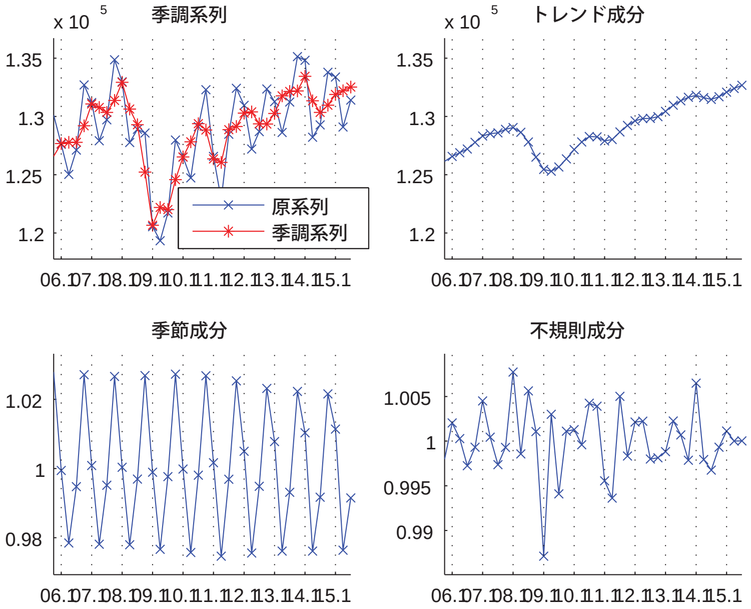
5. 調整結果の比較

■ 実質 GDP(X-11)



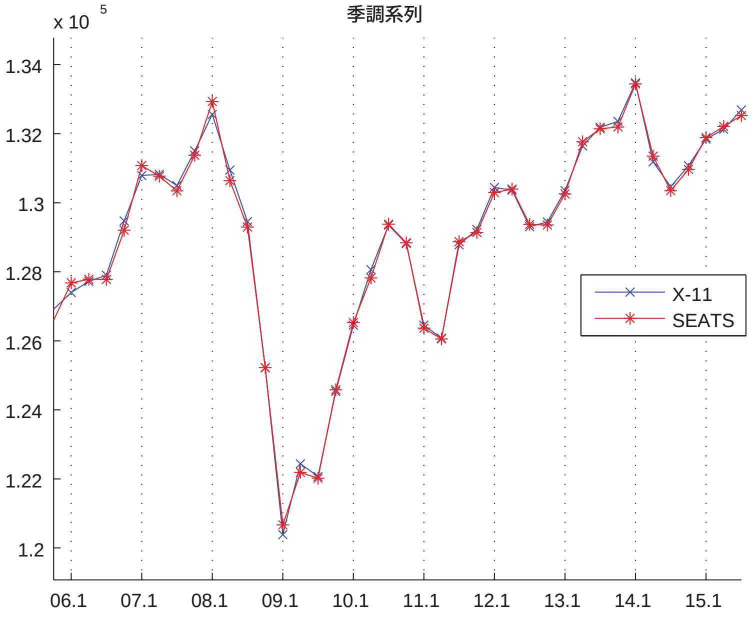
5. 調整結果の比較

■ 実質 GDP(SEATS)



5. 調整結果の比較

■ 実質 GDP(季節調整値)



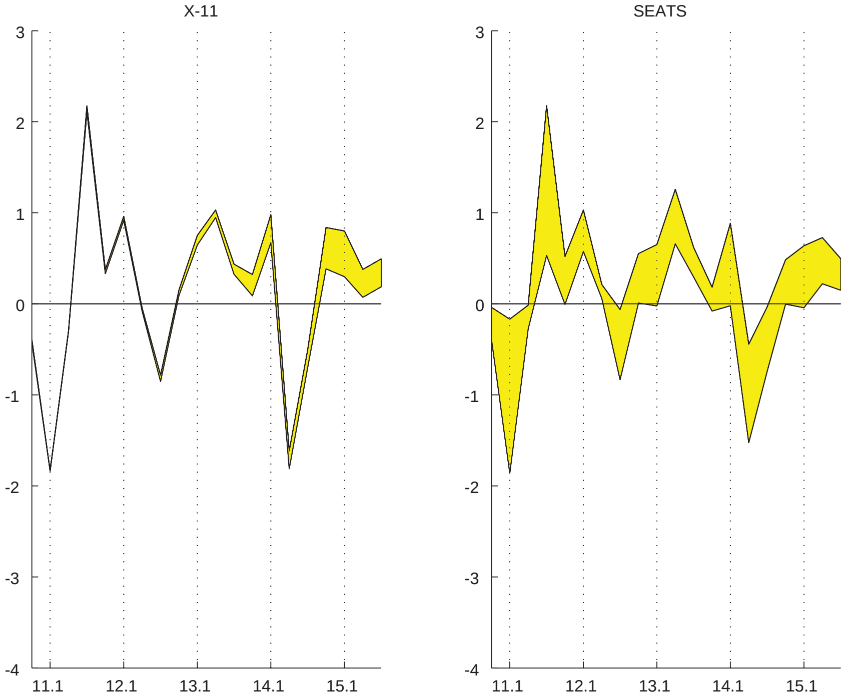
5. 調整結果の比較

■ モデル選択の影響 1

- 四半期別実質 GDP（1994 年 1-3 月〜2015 年 7-9 月）
- automdl コマンドで階差次数を選択
- 階差次数を固定し、SARMA 次数をそれぞれ上限 2 として、X11 コマンドと SEATS コマンドの両ケースについて全てのモデルを推定
- それ以外のスペックはデフォルトで固定
- 各モデルによる季調値から前期比増加率を計算し、各時点での最大値、最小値を計算
- 最大値系列と最小値系列の間を黄色で着色

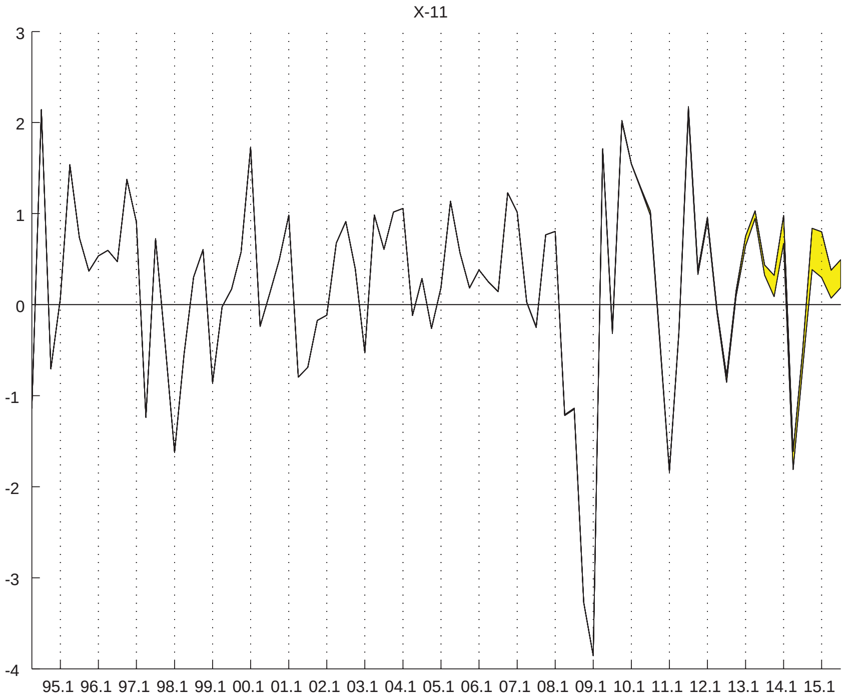
5. 調整結果の比較

■ モデル選択の影響 2



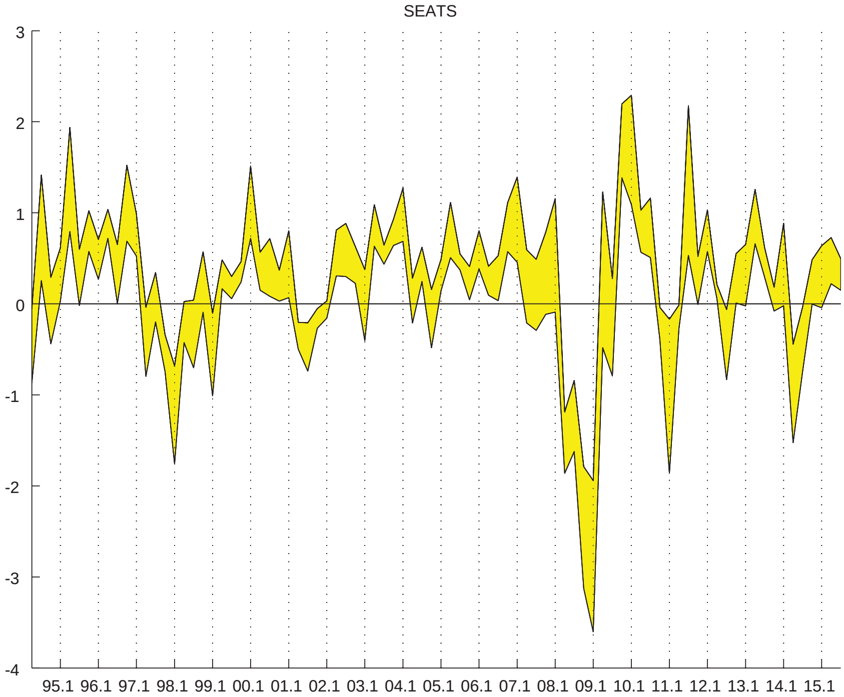
5. 調整結果の比較

■ モデル選択の影響 3



5. 調整結果の比較

■ モデル選択の影響 4



5. 調整結果の比較

■ モデル選択の影響 5

- X-11 ではモデルによる事前調整と平滑化処理が分離しているため、モデル選択結果の過去への影響は小さい。
  ⇒ ただし直近付近では最大 0.5 パーセントポイント程度の幅
- SEATS ではモデルとフィルタが連動しているため、モデルによって過去の調整値も大きく変化する。
  ⇒ 時系列的性質の変化への対応が問題
- automdl を用いた単純な季節調整では、モデル選択法が共通化されたため、X-11 と SEATS の結果差は小さい。
  ⇒ モデルの変更を伴う継続的運用においては、かなりの差異が生じる可能性
- モデル選択が不適切である場合、SEATS では季節調整が不安定になる可能性が高い。

6. まとめ

■ まとめ

- X-12-ARIMA は TRAMO-SEATS と統合される形で X-13ARIMA-SEATS に組み込まれている。
- センサス局では X-12-ARIMA のウエブ上での配布を停止しており、今後は X-13ARIMA-SEATS に情報を発信してゆくと思われる。
- X-11 フィルタと SEATS による WK フィルタの季節調整結果は、確認した範囲では非常に類似した結果を出力する。
- 一方、SEATS による処理ではモデルの変更に伴い、全期間に渡って過去の季調値に改定が生じる可能性がある。