Revealing Unobservables by Deep Learning: Generative Element Extraction Networks (GEEN)

Yingyao Hu* Yang Liu† and Jiaxiong Yao†

October 2, 2023

Abstract

Latent variable models are crucial in scientific research, where a key variable, such as true GDP growth in specific countries or individuals' true earnings, is unobserved in the sample but needs to be identified. This paper proposes a novel method for estimating realizations of a latent variable X^* in a random sample that contains its multiple measurements. With the key assumption that the measurements are independent conditional on X^*, we provide sufficient conditions under which realizations of X^* are locally unique in a class of deviations or are globally point identified. To the best of our knowledge, this paper is the first to provide such identification in observation. We then use the divergence function between the two probability densities with and without the conditional independence as the loss function to train Generative Element Extraction Networks (GEEN) that map from the observed measurements to realizations of X^* in the sample. The simulation results show that this proposed estimator works quite well and the estimated values are highly correlated with realizations of X^*. We then use GEEN to estimate true GDP growth for each developing country using such measurements as official GDP growth, nightlight intensity, and Google search volume. Our estimates show more insightful information on the economies than existing measurements. Given that our estimator can be applied to a large class of latent variable models, we expect it will change how people deal with latent variables.

---

[*] Johns Hopkins University, Department of Economics, Johns Hopkins University, Wyman Park Building 544E, 3400 N. Charles Street, Baltimore, MD 21218, (email: yhu@jhu.edu).

[†] International Monetary Fund, 700 19th St NW, Washington DC 20431 (e-mail: yliu10@imf.org and jyao@imf.org).

1 Introduction

Unobservables play a crucial role in scientific research because empirical researchers often encounter a discrepancy between what is described in a model and what is observed in the data. A typical example is the so-called hidden Markov models, where a series of latent variables are observed with errors in multiple periods under conditional independence assumptions. While there is a huge literature on the estimation of the model with latent variables (e.g.,Aigner, Hsiao, Kapteyn, and Wansbeek (1984); Bishop (1998)), this paper focuses on the estimation of *realizations* of the latent variable, which are not observed anywhere in the data. Suppose that the ideal data for the estimation of a model is an i.i.d. sample of $(X^1, X^2, ..., X^k, X^*)$ [1] and that the researcher only observed $(X^1, X^2, ..., X^k)$ in the sample. Generally, we consider $X^j, j = 1...k$, as multiple measurements of X^*. Under conditional independence assumptions, this paper provides a deep learning method to extract the common element X^* from multiple observables $(X^1, X^2, ..., X^k)$. We build Generative Element Extraction Networks (GEEN) to reveal realizations or draws of X^* to achieve a complete sample of $(X^1, X^2, ..., X^k, X^*)$ in the sense that the generated draws are observationally equivalent to the true values in the sample.

This paper is different from the imputation method because the latent variable is not observed anywhere in the sample and needs to be identified. For example, true earnings of households are not observed anywhere in household survey data, but are of great interest to know. By contrast, imputation requires at least some observations of the underlying variable.

Researchers have already applied deep generative models for data imputation. Yoon, Jordon, and Schaar (2018) creatively use the Generative Adversarial Imputation Nets (GAIN) to provide an imputation method, in which missing values are estimated so that they are observationally equivalent to the observed values from the GAIN's perspective. Li, Jiang, and Marlin (2019) also propose a GAN-based Goodfellow, Pouget-Abadie, Mirza, Xu, Warde-Farley, Ozair, Courville, and Bengio (2014) framework for learning from complex, high-dimensional incomplete data to impute missing data. Mattei and Frellsen (2019) introduce the missing data importance-weighted autoencoder for training a deep latent variable model to handle missing-at-random data. Nazabal, Olmos, Ghahramani, and Valera (2020) present a general framework for Variational Autoencoders (VAEs) Kingma and Welling (2013) that effectively incorporates incomplete data and heterogenous observations. Muzellec, Josse, Boyer, and Cuturi (2020) leverage optimal transport to define a loss function for missing value imputation. Yoon and Sull (2020) propose a novel Generative Adversarial Multiple Imputation Network (GAMIN) for highly missing Data. In this literature, latent spaces are used to represent high-dimensional observations, but are not identifiable because their latent spaces may vary with parameter initialization. In addition, all the missing data models require true values to be partially observed.

---

[1] We use capital letters to stand for a random variable and lower case letters to stand for the realization of a random variable. For example, $f_V(v)$ stands for the probability density function of random variable V with realization argument v, and $f_{V|U}(v|u)$ denote the conditional density of V on U.

However, relatively little research has focused on estimating realizations of latent variables, which are unobserved, or completely missing. In the economics literature, Kalman filter and structural vector autoregressions have often been used to estimate the realizations of latent variables, such as potential output Kuttner (1994), natural rate of interest Laubach and Williams (2003); Holston, Laubach, and Williams (2017), and natural rate of unemployment King and Morley (2007), but the literature makes parametric assumptions about the dynamics of latent variables and thus belongs to the estimation of models with latent variables.

In our setting, we argue that the conditional independence restrictions imply the local identification of the true values. That allows us to provide an estimator in the continuous case. Our method is nonparametric in the sense that we do not assume the distribution of the variables belong to a parametric family as in the widely-used VAEs Kingma and Welling (2013), which use the so-called Evidence Lower Bound (ELBO) to provide a tractable unbiased Monte Carlo estimator. The VAEs focus on the estimation of a parametric model. In this paper, we focus on the estimation of the true values in each observation in the sample without imposing a parametric structure on the distributions.

Our loss function is a distance between two nonparametric density functions with and without the conditional independence. Such a distance is based on a powerful nonparametric identification result in the measurement error literature Hu and Schennach (2008). (See Hu (2017) and Schennach (2020) for a review.) It shows that the joint distribution of a latent variable and its measurements is uniquely determined by the joint distribution of the observed measurements under a key conditional independence assumption, together with other technical restrictions. To measure the distance between two density functions, the Kullback–Leibler divergence Kullback and Leibler (1951) is one of the options, which plays a leading role in machine learning and neuroscience Pérez-Cruz (2008). A large literature has studied the estimation of the Kullback–Leibler divergence Darbellay and Vajda (1999); Moreno, Ho, and Vasconcelos (2003); Wang, Kulkarni, and Verdú (2005); Lee and Park (2006); Wang, Kulkarni, and Verdú (2006); Nguyen, Wainwright, and Jordan (2010); Nowozin, Cseke, and Tomioka (2016); Belghazi, Baratin, Rajeshwar, Ozair, Bengio, Courville, and Hjelm (2018). We use a combination of a deep neural network and kernel density estimators to generate density functions with and without the conditional independence and then compute their divergence.

In this paper, we make a further argument that the nonparametric identification of the latent variable distribution implies that realizations of the latent variable are locally separable in the continuous case. Furthermore, we show that realizations of the latent variable can be globally point identified under an intuitive assumption that no two subjects in the population are observationally identical. To the best of our knowledge, this paper is the first to provide such identification in observation. We expect such identification will change how researchers deal with latent variables and make our GEEN broadly applicable.

This paper is organized as follows. Section 2 provides the identification arguments. Section 3 describe the neural network and the algorithm. The Monte Carlo simulations are provided in Section 4. Section 5 presents an application to refine GDP measurements from official data. Section 6 summarizes the paper.

2 From identification in distribution to identification in observation

We assume that a researcher observe the distribution of $\{X^1, X^2, ..., X^k\}$ from a random sample. Putting the estimation of the population distribution $f_{X^1,X^2,...,X^k}$ from the random sample aside, we face a key identification challenge: How to determine the distribution $f_{X^1,X^2,...,X^k,X^*}$ from the observed distribution $f_{X^1,X^2,...,X^k}$. Here we use a general nonparametric identification result in the measurement error literature.

Theorem 2 *Hu and Schennach (2008) Under assumptions 2, 3, 4, 5, and 6 in the Appendix, the joint distribution $f_{X^1,X^2,...,X^k}$ uniquely determines the joint distribution $f_{X^1,X^2,...,X^k,X^*}$, which satisfies*

$$f_{X^1,X^2,...,X^k,X^*} = f_{X^1|X^*} f_{X^2|X^*} \cdot ... \cdot f_{X^k|X^*} f_{X^*}. \tag{1}$$

This identification result only needs three measurements. Therefore, the conditional independence may be relaxed to

$$f_{X^1,X^2,...,X^k,X^*} = f_{X^1|X^*} f_{X^2|X^*} f_{X^3,...,X^k,X^*}.$$

In the remaining discussion, we still use the conditional independence in equation (1) because we are interested in the common element X^* across all the observables. This identification result implies that if we have qualified measurements X^1, X^2 and X^3, we are able to provide a consistent estimator of $f_{X^1,X^2,...,X^k,X^*}$ from a sample of $(X^1, X^2, ..., X^k)$.

2.1 Uncorrelated deviations and identification in observation

Next, we argue that draws of X^* are locally identified in the sense that there is no observationally equivalent uncorrelated deviation from these draws.

Let X_i^* be a random draw of X^* in observation i and we define *an uncorrelated deviation* from that draw as

$$X_i^* + \delta_i \quad \text{with} \quad E(X_i^* \delta_i) = E(\delta_i) = 0 \tag{2}$$

where (X_i^*, δ_i) is a i.i.d. random draw from the joint distribution of (X^*, δ). Notice that if we replace X_i^* with $X_i^* + \delta_i$ as the new common element, the variance of the common element becomes $var(X^*) + var(\delta)$. That means the variance of the uncorrelated deviation must be different from that of the original X^*, i.e., $var(X^*)$. The distribution of $X^* + \delta$ must be different from that of X^*. These two different distributions can not lead to the same observed distribution, $f_{X^1,X^2,...,X^k}$, because Theorem 2 implies that $f_{X^1,X^2,...,X^k}$ uniquely determines f_{X^*}, including its variance $var(X^*)$. In other words, $X^* + \delta$ and X^* can not be observationally equivalent given the sample of $(X^1, X^2, ..., X^k)$. Therefore, the draws of X^* are locally identified in the following sense:

Theorem 3 *Suppose that the assumptions in Theorem 2 hold. Given an observed sample* $\{X_i^1, X_i^2, ..., X_i^k\}$*, which is a subset of the infeasible full sample* $\{X_i^1, X_i^2, ..., X_i^k, X_i^*\}$*, no uncorrelated deviation from latent draws* X_i^**, defined in equation (2), is observationally equivalent to* X_i^**.*

Notice that we only use the identified variance of the latent X^* to make this argument. The results in Theorem 2 implies that all the moments of the latent X^* are identified. Therefore, such a local identification result as in Theorem 6 should hold for more general deviations than the uncorrelated deviations defined in equation (2).

Furthermore, we may look at this problem from a different angle. Suppose we insert generated draws $\hat{X}_i^*$ in the sample $\{X_i^1, X_i^2, ..., X_i^k\}$ to obtain $\{X_i^1, X_i^2, ..., X_i^k, \hat{X}_i^*\}$. And we also suppose that the conditional independence in equation (1) holds with the generated draws, i.e.,

$$f_{X^1,X^2,...,X^k,\hat{X}^*} = f_{X^1|\hat{X}^*} f_{X^2|\hat{X}^*} \times ... \times f_{X^k|\hat{X}^* z} f_{\hat{X}^*}.$$

In this case, even if $\hat{X}_i^*$ is not equal to the true X_i^* in the infeasible full sample $\{X_i^1, X_i^2, ..., X_i^k, X_i^*\}$, our inserted $\hat{X}_i^*$ will be observationally equivalent to the true X_i^* because Theorem 2 guarantees that the distributions $f_{X^1|X^*}$, $f_{X^2|X^*}$, and $f_{X^3,...,X^k,X^*}$ are uniquely determined by the observed $f_{X^1,X^2,...,X^k}$. Even if $\hat{X}_i^* \neq X_i^*$, we can still correctly estimate $f_{X^1|X^*}$, $f_{X^2|X^*}$, and $f_{X^3,...,X^k,X^*}$ using sample $(X_i^1, X_i^2, ..., X_i^k, \hat{X}_i^*)_{i=1,2,...,N}$ with inserted $\hat{X}_i^*$, instead of the true values X_i^*.

In addition, Theorem 6 implies that if we add a noise δ_i to the inserted $\hat{X}_i^*$, where δ_i is an uncorrelated deviation from $\hat{X}_i^*$, the conditional independence fails when $\hat{X}_i^*$ is replaced with $\hat{X}_i^* + \delta_i$. That means the inserted draws $\hat{X}_i^*$ are locally unique among uncorrelated deviations.

The identification result in Theorem 6 can be extended to the case where δ_i is uncorrelated with X_i^* conditional on the observables $(X^1, X^2, ..., X^k)$ because the conditional distribution $f_{X^*|X^1,X^2,...,X^k}$ is identified by Theorem 2. We define *a conditionally uncorrelated deviation* from X_i^* as $X_i^* + \delta_i$ with

$$E(X_i^* \delta_i \,|\, X_i^1, X_i^2, ..., X_i^k) = E(\delta_i \,|\, X_i^1, X_i^2, ..., X_i^k) = 0 \tag{3}$$

where $(X_i^*, \delta_i, X_i^1, X_i^2, ..., X_i^k)$ is a i.i.d. random draw from their corresponding joint distribution. The variance, and therefore distribution, of $X_i^* + \delta_i$ conditional on $(X^1, X^2, ..., X^k)$ is different from those of $f_{X^*|X^1,X^2,...,X^k}$. Theorem 2 implies that they must correspond to different $f_{X^1,X^2,...,X^k}$. Therefore, there is no observationally equivalent conditionally uncorrelated deviation from latent draws X_i^*. We summarize this extension as follows:

Theorem 4 *Suppose that the assumptions in Theorem 2 hold. Given an observed sample* $\{X_i^1, X_i^2, ..., X_i^k\}$*, which is a subset of the infeasible full sample* $\{X_i^1, X_i^2, ..., X_i^k, X_i^*\}$*, no conditionally uncorrelated deviation from latent draws* X_i^**, defined in equation (3), is observationally equivalent to* X_i^**.*

Suppose our identification results suggest that our estimates $\hat{X}_i^*$ should have the same distribution (and variance) as X_i^*. Then the sample moments of $\hat{X}_i^*$ should converge to the true moments. We may make a convergence argument as follows:

Theorem 5 *Suppose that the estimator* $\hat{X}_i^* = X_i^* + \delta_i$ *for* $i = 1, 2, ..., N$ *satisfies*

$$\frac{1}{N} \sum_{i=1}^{N} X_i^* \delta_i = o_p(1). \tag{4}$$

Then, the consistency of the sample variance of $\hat{X}_i^*$ implies that for any $\epsilon > 0$, the sample proportion of large deviations goes to zero, i.e.,

$$P_N\left(\left|\hat{X}_i^* - X_i^*\right| > \epsilon\right) := \frac{1}{N}\sum_{i=1}^{N} I(|\delta_i| > \epsilon) = o_p(1).$$

We leave the details in the Appendix.

2.2 Point identification in observation

The identification with respect to uncorrelated deviations is still not enough to pin down the value of the unobserved in each observation. Here we provide a point identification argument for the value of the unobserved variable in each observation. Let $X_i = (X_i^1, X_i^2, ..., X_i^k)^T$ and X_i^* be observed and unobserved variables describing subject i, respectively. Define the population of random variables $\{X_i, X_i^*\}$ as set

$$\mathcal{P}_{X,X^*} = \{(x_i, x_i^*) : i = 1, 2, ..., \infty\}.$$

Suppose that one never observes two subjects alike. Then the mapping from observed X to unobserved X^* is a function, i.e., a set of ordered pairs (x, x^*) in which no two different ordered pairs have the same first coordinate. Therefore, such a function can uniquely determine unobserved x^* from observed x for each subject. Formally, we assume

Assumption 1 *No two different subjects in the population are observationally identical, i.e., for any (x_i, x_i^*) and (x_j, x_j^*) in $\mathcal{P}_{X,X^*}$, $i \neq j$ implies $x_i \neq x_j$.*

Under this property, the mapping defined by set $\mathcal{P}_{X,X^*}$, i.e., the population of $\{X_i, X_i^*\}$, is a function mapping from X to X^*. Notice that x_i^* is not necessarily distinct for different subject i.

Define such a function as

$$X^* = G(X).$$

Then the identification of distribution $f(X, X^*)$ implies the identification of function G. In fact, we have

$$G(X) = E[X^*|X]$$

In fact, the property in Assumption 1 implies that the distribution of X^* degenerates conditional on X. Therefore, any location, such as conditional mean or median, of the distribution $f(X^*|X)$ can reveal the value of x_i^* in each observation i. Therefore, we can extend the identification in distribution in in Theorem 2 as follows:

Theorem 6 *Suppose that the assumptions in Theorem 2 and Assumption 1 hold. Given an observed sample* $\{X_i^1, X_i^2, ..., X_i^k\}$*, which is a subset of the infeasible full sample* $\{X_i^1, X_i^2, ..., X_i^k, X_i^*\}$*, the realization of* X_i^* *in observation* i *is uniquely determined by the realization of* $\{X_i^1, X_i^2, ..., X_i^k\}$ *in the observation. In particular,*

$$x_i^* = E[X^*|X_i^1 = x_i^1, X_i^2 = x_i^2, ..., X_i^k = x_i^k].$$

In most economic applications, Assumption 1 is satisfied because we usually do not have two observations which are exactly the same. In the case where this happens, the point identification result here implies that researchers need to obtain more measurement or observables so that identification of unobservable is possible. Here we provide an example to illustrate the identification.

Suppose we have for $j = 1, 2, 3$

$$\begin{aligned} X^j &= m^j(X^*) + \epsilon^j \\ X^4 &= exp(X^*) \times (-1)^{I(\epsilon^4 > 0.5)} + m^4(X^3.) \end{aligned}$$

and (X^1, X^2, X^3) satisfies the assumptions in Theorem 2. In particular, we assume $(\epsilon^1, \epsilon^2, \epsilon^3, \epsilon^4, X^*)$ are jointly independent. So distribution $f(X^1, X^2, X^3)$ uniquely determines distribution $f(X^1, X^2, X^3, X^*)$. Assumption 1 is satisfied because

$$X^* = ln|X^4 - m^4(X^3)|$$

The value of $(\epsilon^1, \epsilon^2, \epsilon^3)$ can be determined by a given value of (X^1, X^2, X^3, X^*). When ϵ^4 is a 0-1 binary random variable, the value of ϵ^4 can also be determined by a given value of (X^2, X^3, X^4, X^*).

Notice that

$$f(X^1, X^2, X^3, X^4) = \int f(X^1|X^*) f(X^*, X^2, X^3, X^4) dX^*.$$

Given that $f(X^1|X^*)$ is identified from $f(X^1, X^2, X^3)$ and its corresponding integral operator is injective, distribution $f(X^*, X^2, X^3, X^4)$ is also identified. Therefore, the whole distribution $f(X^1, X^2, X^3, X^4, X^*)$. In fact, one can show that

$$E[X^*|X^1, X^2, X^3, X^4] = ln|X^4 - m^4(X^3)|.$$

Therefore, the realization of X^* is uniquely determined and also estimable in each observation.

3 Generative Element Extraction Networks (GEEN)

The discussion above implies that we can use a loss function measuring the distance between a general joint distribution $p = f_{X^1,X^2,...,X^k,X^*}$ and a distribution satisfying conditional independence $p_{ci} = f_{X^1|X^*}f_{X^2|X^*}...f_{X^k|X^*}f_{X^*}$ in order to search for latent draws X_i^*. One of the choices is the Kullback–Leibler divergence

$$D_{KL}\left(p(x)||p_{ci}(x)\right) = \int p(x)\ln\left(\frac{p(x)}{p_{ci}(x)}\right)dx.$$

We build a Generative Element Extraction Network (GEEN), G, to generate the latent realizations of X_i^* satisfying the conditional independence. Let $\vec{V}$ stand for the vector of draws of variable V in the sample, i.e., $\vec{X}^* = (X_1^*, X_2^*, ..., X_N^*)^T$ and $\vec{X}^j = (X_1^j, X_2^j, ..., X_N^j)^T$. We generate $\vec{\hat{X}}^*$ as follows:

$$\vec{\hat{X}}^* = G(\vec{X}^1, \vec{X}^2, ..., \vec{X}^k). \tag{5}$$

with $\vec{\hat{X}}^* = (\hat{X}_1^*, \hat{X}_2^*, ..., \hat{X}_N^*)^T$. The deep neural network G is trained to minimize the divergence

$$\min_G D\left(\hat{p}\,,\,\hat{p}_{ci}\right) \quad s.t. \int x\hat{f}_{X^1|\hat{X}^*}(x|x^*)dx = x^* \tag{6}$$

with $\hat{p} = \hat{f}_{X^1,X^2,...,X^k,\hat{X}^*}$ and $\hat{p}_{ci} = \hat{f}_{X^1|\hat{X}^*}\hat{f}_{X^2|\hat{X}^*}...\hat{f}_{X^k|\hat{X}^*}\hat{f}_{\hat{X}^*}$, where $\hat{f}$ are empirical distribution functions based on sample $(\vec{X}^1, \vec{X}^2, ..., \vec{X}^k, \vec{\hat{X}}^*)$.

Notice that G enters the loss function through $\vec{\hat{X}}^* = (\hat{X}_1^*, \hat{X}_2^*, ..., \hat{X}_N^*)^T$ in density estimators. To be specific, we can have a kernel density estimator

$$\hat{f}_{X^j|\hat{X}^*}(x|x^*) = \frac{\hat{f}_{X^j,\hat{X}^*}(x,x^*)}{\hat{f}_{\hat{X}^*}(x^*)}$$

$$\hat{f}_{X^j,\hat{X}^*}(x,x^*) = \frac{1}{m}\sum_{i=1}^{m}\frac{K((X_i^j - x)/h^j)}{h^j}\frac{K((\hat{X}_i^* - x^*)/h^*)}{h^*}$$

$$\hat{f}_{\hat{X}^*}(x^*) = \frac{1}{m}\sum_{i=1}^{m}\frac{K((\hat{X}_i^* - x^*)/h^*)}{h^*}$$

$$\begin{aligned}&\hat{f}_{X^1,X^2,...,X^k,\hat{X}^*}(x^1,x^2,...,x^k,x^*)\\ =\;& \frac{1}{m}\sum_{i=1}^{m}\left(\frac{K((\hat{X}_i^* - x^*)/h^*)}{h^*}\prod_{j=1}^{k}\frac{K((X_i^j - x^j)/h^j)}{h^j}\right)\end{aligned}$$

where h stands for bandwidths, N is the total sampled observations, m is the number of points in each observation and k is the number of features. In the loss function defined in equation (6), it requires more than one data point to estimate the kernel density function. As a result, unlike other use cases that one training point is enough to calculate its corresponding loss, we need to sample m (> 1) points as one observation to calculate its loss. For example, to build the training sample we sample with replacement m points from the entire training data points and repeat N times, and we end up with N observations in our training sample. The same practice is followed to construct our validation and test samples. The kernel function $K(\cdot)$ can simply be the standard normal density function. For the bandwidth, we adopt the so-called Silverman's rule, i.e., $h^j = w\sigma^j N^{-1/5}$ where σ^j is the standard error of X^j, and w is the window size that is determined by hyper parameters tuning. Similarly, we may take $h^* = w\sigma^* N^{-1/5}$, where σ^* is the standard error of X^*.

In this paper, we experiment GEEN with multilayer perceptrons (MLPs), but this framework can be readily applied to other deep neural network architectures. In our simulations, we impose a convolution structure on X^1 so that the normalization condition can be simplified. The parameters of our deep neural network are estimated by minimizing the loss function:

$$\text{Loss} = D\left(\hat{p}\,,\,\hat{p}_{ci}\right) + \lambda\left|\frac{1}{N}\sum_{i=1}^{N} X_i^1 - \frac{1}{N}\sum_{i=1}^{N}\hat{X}_i^*\right|^2$$

Early stopping is applied when the loss does not improve for certain epochs in the validation sample. We do not use any information from true X_i^* during training, validation or hyper parameters tuning. Instead we use the loss defined in the above equation for validation and true X_i^* are only used for final testing.

4 Simulations

This section presents the performance of our neural network through simulations. We generate the sample as follows:

$$X_i^j = m^j(X_i^*) + \epsilon_i^j \tag{7}$$

for $j = 1, 2, ..., k$ and $i = 1, 2, ..., N$. Without loss of generality, we normalize $m^1(x) = x$ and $E[\epsilon^1|X^*] = 0$. We pick distributions for $(\epsilon^1, ..., \epsilon^k, X^*)$ and functions $(m^2, ..., m^k)$ to generate a sample $(X^1, ..., X^k, X^*)$. We then train G using the observed sample $(\vec{X}^1, \vec{X}^2, ..., \vec{X}^k)$ to generate $(\vec{X}^1, \vec{X}^2, ..., \vec{X}^k, \hat{X}^*)$. That is $\vec{\hat{X}}^* = G(\vec{X}^1, \vec{X}^2, ..., \vec{X}^k)$. We check the performance of G by calculating the correlation coefficient between $\vec{X}^*$ and $\vec{\hat{X}}^*$.

We use a 6-layer with 10 hidden nodes fully connected neural network. The window size w and normalization term λ are tuned as hyper-parameters. We use kernel functions to approximate their density functions. Theoretically if a distribution is normal, the best choice for w used in the kernel function is 1, so to tune w we choose the range from 0.5 to 2. To tune λ, we arbitrarily choose the range from 0.1 to 0.5. For every experiment, we run 25 times to evaluate the robustness of model performance on its initialization. For the baseline case, we use

$$\begin{aligned} k &= 4 & \epsilon^1 &= N(0,1) \\ m^1(x) &= x & \epsilon^2 &= Beta(2,2) - \frac{1}{2} \\ m^2(x) &= \frac{1}{1+e^x} & \epsilon^3 &= Laplace(0,1) \\ m^3(x) &= x^2 & \epsilon^4 &= Uniform(0,1) - \frac{1}{2} \\ m^4(x) &= \ln(1+e^x) & X^* &= N(0,4) \end{aligned}$$

We sample 8000 points as training points from the above distributions for X^*, ϵ^1, ϵ^2, ϵ^3 and ϵ^4. Then we sample another 1000 points for validation points and 1000 points for test points. We draw 500 points from the training points with replacement 8000 times to build our training set and 1000 times from the validation/test points to build our validation/test set. Figure 1 shows the relationship between X^1, X^2, X^3, X^4 and X^*.

In the second experiment, we let the error terms correlate with X^* while keeping the rest setup the same as the baseline. Figure 2 shows the relationship between X^1, X^2, X^3, X^4 and X^* in this setup. Specifically, we use:

$$\begin{aligned} \epsilon^1 &= N(0, \frac{1}{4}(x^*)^2) & \epsilon^3 &= Laplace(0, \frac{1}{2}|x^*|) \\ \epsilon^4 &= Uniform(0, \frac{1}{2}|x^*|) - \frac{1}{4}|x^*| \end{aligned}$$

In the third experiment, we double the variance of the error terms while keeping the rest setup the same as the baseline. Figure 3 shows the relationship between X^1, X^2, X^3, X^4 and X^* in the third experiment.

$$\begin{aligned} \epsilon^1 &= N(0,4) & \epsilon^3 &= Laplace(0,2) \\ \epsilon^2 &= Beta(2,4) - \frac{1}{3} & \epsilon^4 &= Uniform(0,2) - 1 \end{aligned}$$

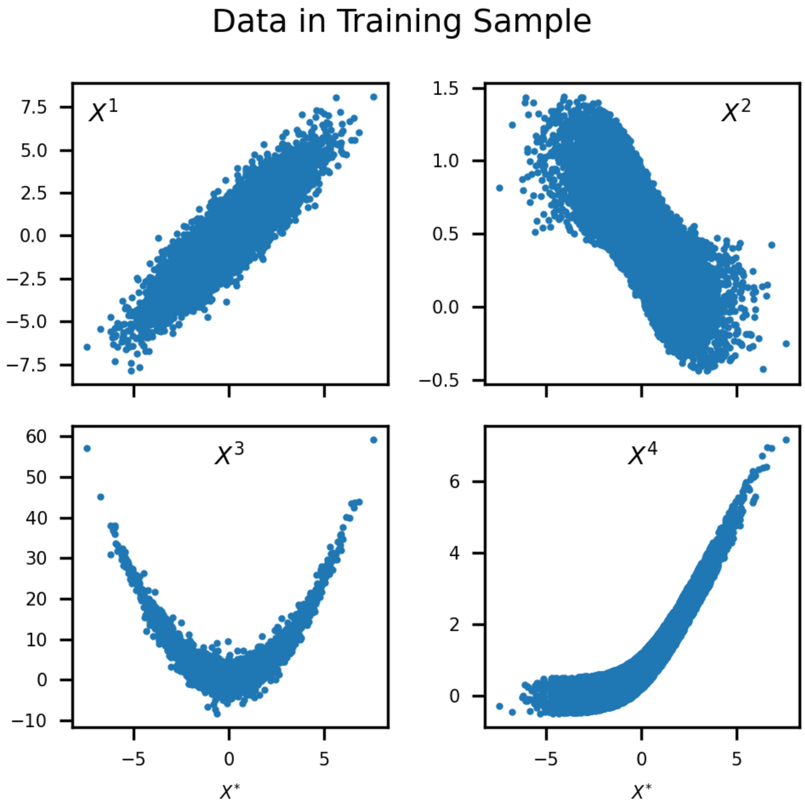


Figure 1: Baseline Training Sample

Table 1 demonstrates the min, median and max correlations of $\vec{X}^*$ and $\hat{\vec{X}}^*$ in the test sample for the three experiments after running each one 25 times. Figure 4 shows their best runs respectively. It shows that GEEN is robust with randomly picked initial values of the parameters and provides a better measurement of X^* than X^1. Figure 5 demonstrates that the correlation between X^* and the generated X^* is well above 0.9 for the baseline and the linear error case and remains strong when the variance is doubled in the thrid experiment.

In the forth experiment, we loosen the normalization condition while keeping the rest setup the same as the baseline. Figure 6 shows the relationship between X^1, X^2, X^3, X^4 and X^* in this experiment.

$$m^1(x) \quad = \quad x^2 + x$$

With this setup, generated $\hat{X}^*$ is not anchored, and as shown in the left hand side of Figure 7 its values deviate significantly from X^*. However, the KL loss helps keep the similarity of the two distributions of generated $\hat{X}^*$ and X^*. As shown in the right hand side of Figure 7, with 25 runs of this experiment most of the absolute values of the correlation between $\vec{X}^*$ and $\hat{\vec{X}}^*$ are around 0.9. This suggests that even without

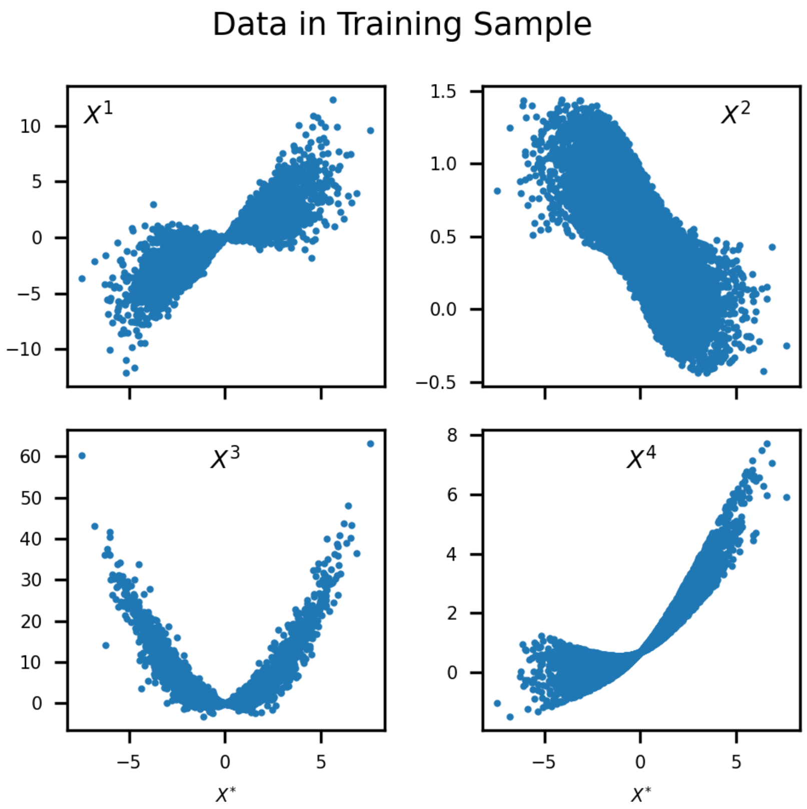


Figure 2: Linear Error Training Sample

normalization our framework can still help provide an estimation of the direction.

5 Refining official GDP measurements

One of the important applications of our methodology is to reduce measurement errors. In this case, true values are unobservables X^*. X^1 is a direct measure of X^* with the expected measurement error δ as zero. X^j $(j \neq 1)$ are indirect/direct measures of X^* with unknown function forms of X^*. Their error terms can be flexible and do not necessarily have zero means.

We apply GEEN to refine GDP data using official GDP data (X^1) and alternative measures of economic activity, including satellite-recorded nighttime lights Hu and Yao (2022) and Google Search Volume Woloszko (2021) as X^2 and X^3. In this experiment, true GDP (X^*) are completely unknown. With GEEN, we demonstrate how our method can help reduce measurement errors from official GDP data.

Our sample consists of all the developing countries that have quarterly GDP data. We focus on developing countries because nighttime lights data are more appropriate for tracking economic activity in those countries (Hu and Yao, 2022; Beyer, Hu, and

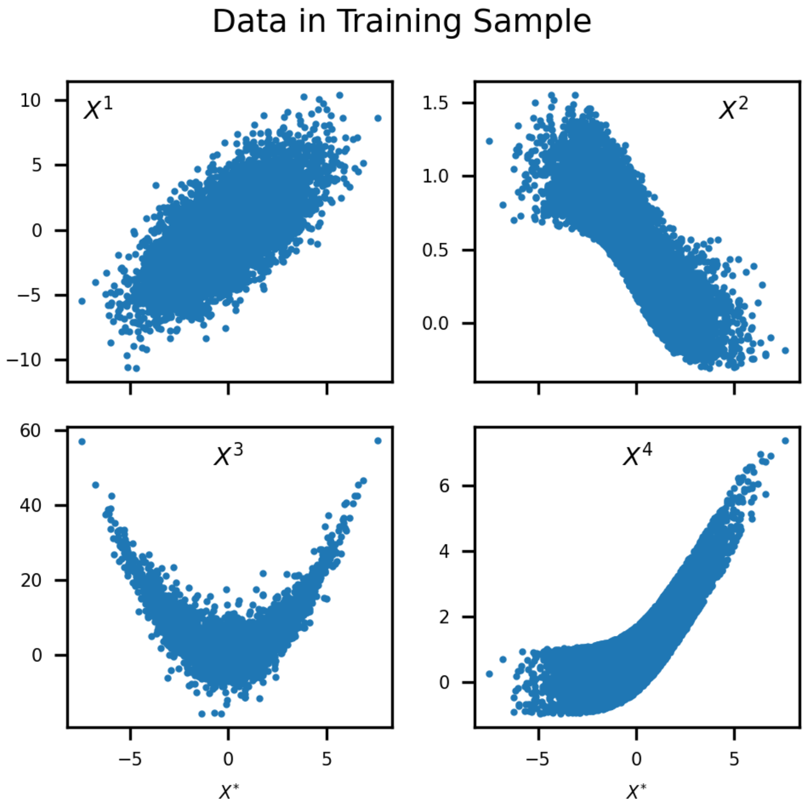


Figure 3: Double Error Training Sample

Yao, 2022). To account for time trends common to all countries, we remove time effects from GDP growth rates with a fixed effect model when training and later add back the time effects when comparing our model's performance with official data. We separate our sample into training and validation subsets, and run training for 100 times and select the best model with the lowest loss in the validation sample to minimize the impact of initialization. We do not have the testing dataset, since in this case true values X^* are completely unknown and the model just learns how to generate X^* that can minimize the distance between the two probability densities in equation (1). Therefore, conventional testing method is not applicable here. Instead, we compare our generated GDP growth rates with official data from the macroeconomic viewpoint, which is crucial to reveal systematic differences between official data and true underlying GDP growth data.

In Figure 8, the left axis is GDP growth rate in percentage points (ppts) and the right axis marks the difference between the official GDP and our generated underlying GDP growth rates (Official - GEEN as shown in the plot). Figure 8 shows that refined GDP data reveal important patterns in official GDP data and are useful in a number of aspects. First, most countries' official GDP growth data align well with our refined

Table 1: Summary of Simulation Results

Simulation Name	$\text{corr}(\vec{X}^*, \hat{\vec{X}}^*)$			$\text{corr}(\vec{X}^*, \vec{X}^1)$
	min	median	max	
Baseline	0.97	0.98	0.98	0.89
Linear Error	0.93	0.96	0.97	0.89
Double Error	0.80	0.89	0.91	0.70

estimates. For example, both Chile and South Africa have differences within 0.15 percentage points despite volatile economic growth. It suggests that GEEN could be useful in leveraging alternative data to understand economic activity of countries without timely official GDP data.

Second, some countries, such as China and Indonesia, have excessively smooth official GDP data compared to our refined estimates. Such excess smoothness might mask underlying dynamics and volatility of economic activity (for countries like Indonesia and China, an adjustment of 0.5 percentage points in GDP is considered significant). Estimates of underlying economic growth could therefore enrich policymakers' understanding of the state of macroeconomy, including output gap and inflationary pressures, and inform efficient policy making.

Third, some economies' official GDP growth data systematically differ from our refined data. For example, when Lebanon's economy shrank after 2017, official data systematically overstated the performance of the economy. By contrast, Jordan's official data systematically understated economic growth. A plausible explanation is the existence of the informal sector that official data do not capture. However, detecting such difference is an important first step in exploring the reasons behind it, be it capacity of the statistical agency, recording of the informal sector, or the political economy.

Results of the refined GDP estimates for the rest countries can be found in the Appendix.

6 Conclusion

This paper uses Generative Element Extraction Networks (GEEN) to reveal unobserved variables in a random sample, which contains multiple measurements of a latent variable of interest. We present the sufficient conditions, under which the joint distribution of a latent variable and its measurements can be uniquely determined. We

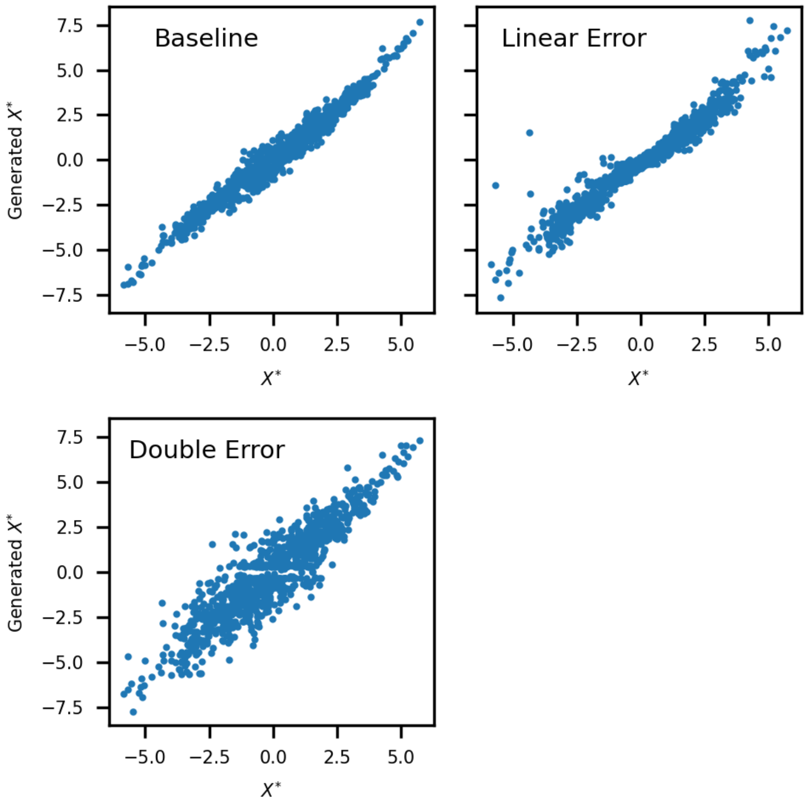


Figure 4: Results in the First Three Experiments

then argue that the true values of the latent variable in the sample are locally unique in a class of deviations, which allows us to estimate the true values. Furthermore, we provide an argument to show that realizations of a latent variable can be globally point identified. To the best of our knowledge, this paper is the first to provide such identification in observation. Based on the key assumption that the measurements are independent conditional on the latent variable, we then propose an algorithm to minimize the divergence function between two probability densities with and without the conditional independence to train GEEN, which maps from the observed measurements to the true values of the latent variable in the sample. The simulation results show that this proposed estimator works quite well and the estimated values are highly correlated with the true values with a correlation coefficient usually higher than 90%. We then use GEEN to estimate true GDP growth for each developing country using its official GDP growth, nightlight intensity, and Google search volume. For different types of economies, our estimates show more meaningful and insightful information than official GDP growth. We expect the GEEN estimator will change how researchers deal with latent variables in empirical research.

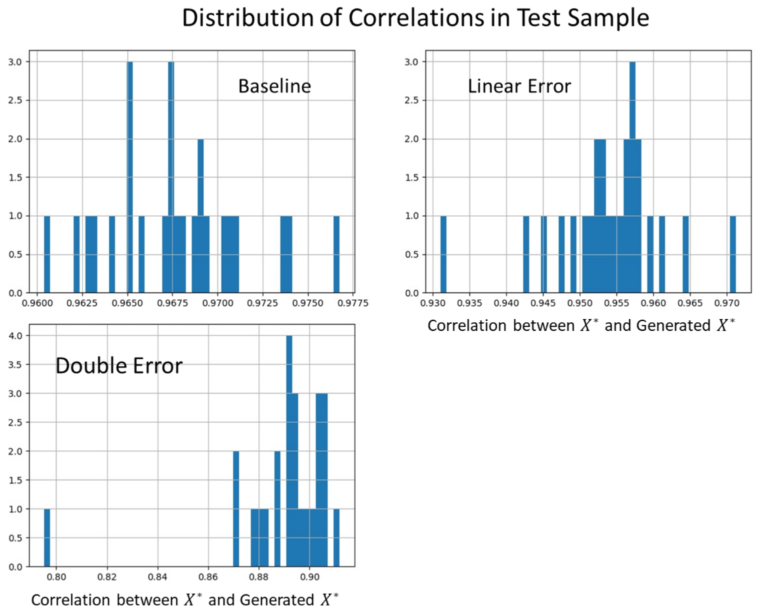


Figure 5: Results in the First Three Experiments

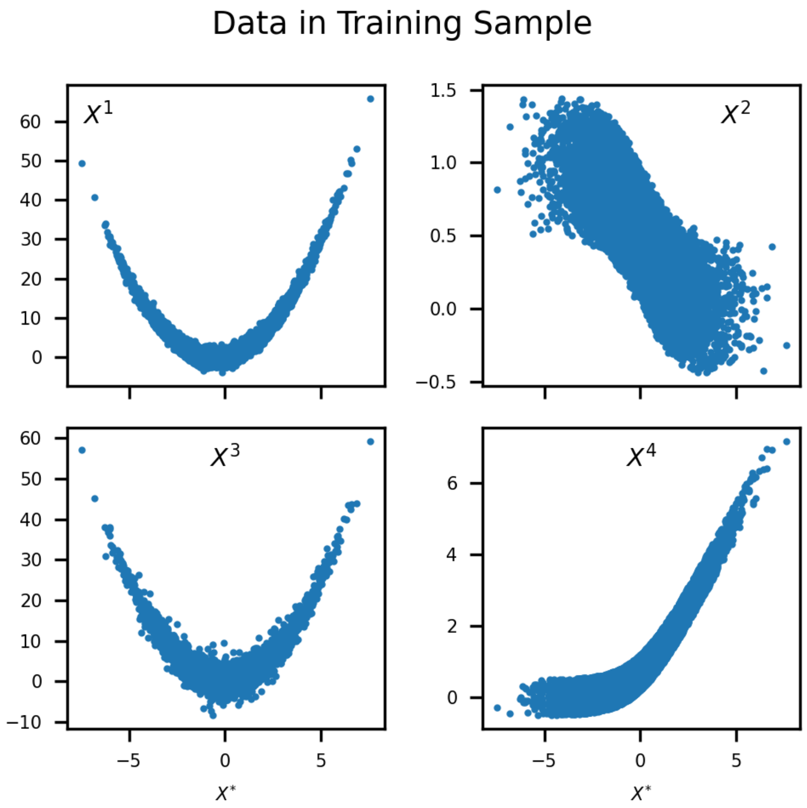


Figure 6: No Normalization Training Sample

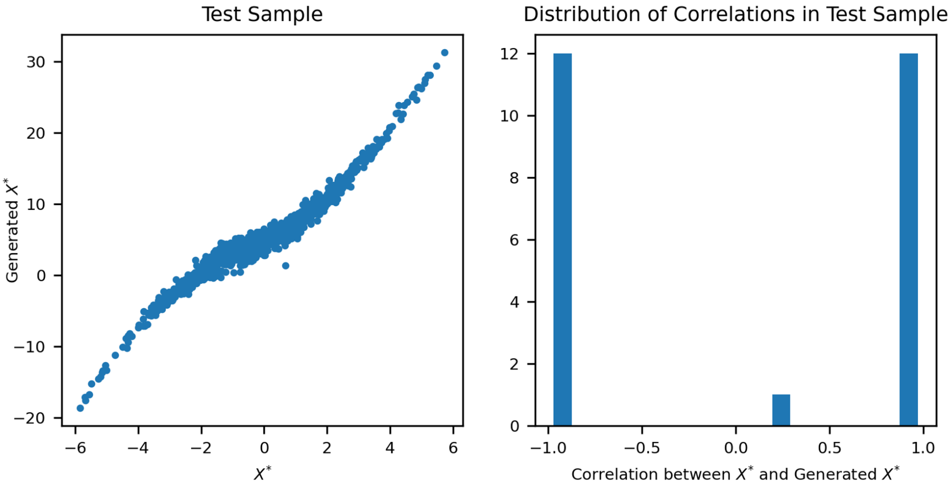


Figure 7: Results in No Normalization Experiment

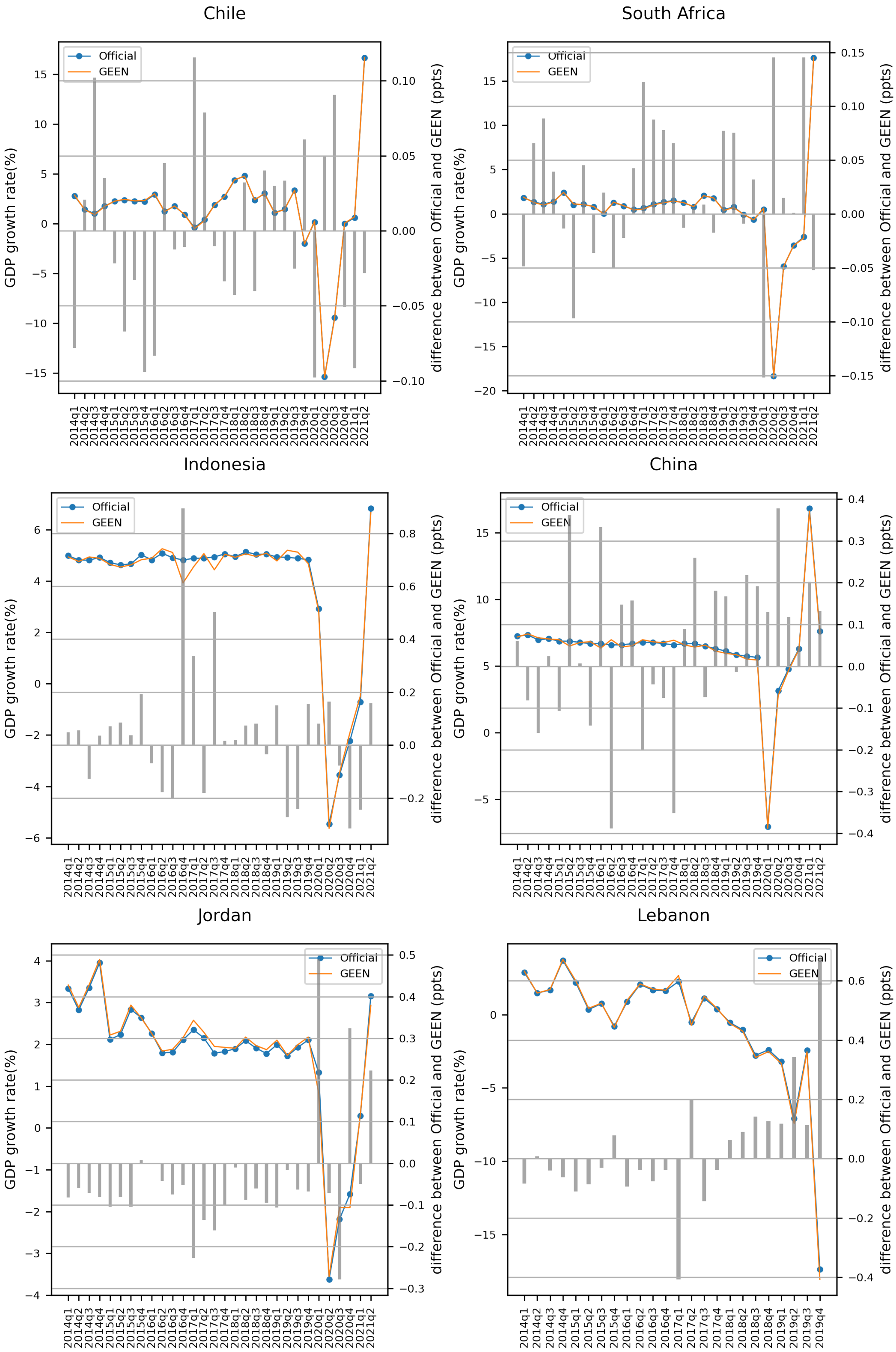


Figure 8: Country Examples of Official and GEEN-refined GDP Growth

References

Aigner, D. J., C. Hsiao, A. Kapteyn, and T. Wansbeek (1984): "Latent variable models in econometrics," *Handbook of econometrics*, 2, 1321–1393.

Belghazi, M. I., A. Baratin, S. Rajeshwar, S. Ozair, Y. Bengio, A. Courville, and D. Hjelm (2018): "Mutual information neural estimation," in *International conference on machine learning*, pp. 531–540. PMLR.

Beyer, R., Y. Hu, and J. Yao (2022): "Measuring Quarterly Economic Growth from Outer Space," .

Bishop, C. M. (1998): "Latent variable models," in *Learning in graphical models*, pp. 371–403. Springer.

Darbellay, G. A., and I. Vajda (1999): "Estimation of the information by an adaptive partitioning of the observation space," *IEEE Transactions on Information Theory*, 45(4), 1315–1321.

Goodfellow, I., J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio (2014): "Generative adversarial nets," *Advances in neural information processing systems*, 27.

Holston, K., T. Laubach, and J. C. Williams (2017): "Measuring the natural rate of interest: International trends and determinants," *Journal of International Economics*, 108, S59–S75.

Hu, Y. (2017): "The econometrics of unobservables: Applications of measurement error models in empirical industrial organization and labor economics," *Journal of econometrics*, 200(2), 154–168.

Hu, Y., and S. M. Schennach (2008): "Instrumental variable treatment of nonclassical measurement error models," *Econometrica*, 76(1), 195–216.

Hu, Y., and J. Yao (2022): "Illuminating economic growth," *Journal of Econometrics*, 228(2), 359–378.

King, T. B., and J. Morley (2007): "In search of the natural rate of unemployment," *Journal of Monetary Economics*, 54(2), 550–564.

Kingma, D. P., and M. Welling (2013): "Auto-encoding variational bayes," *arXiv preprint arXiv:1312.6114*.

KULLBACK, S., AND R. A. LEIBLER (1951): “On information and sufficiency,” *The annals of mathematical statistics*, 22(1), 79–86.

KUTTNER, K. N. (1994): “Estimating potential output as a latent variable,” *Journal of business & economic statistics*, 12(3), 361–368.

LAUBACH, T., AND J. C. WILLIAMS (2003): “Measuring the natural rate of interest,” *Review of Economics and Statistics*, 85(4), 1063–1070.

LEE, Y. K., AND B. U. PARK (2006): “Estimation of Kullback–Leibler divergence by local likelihood,” *Annals of the Institute of Statistical Mathematics*, 58(2), 327–340.

LI, S. C.-X., B. JIANG, AND B. MARLIN (2019): “Misgan: Learning from incomplete data with generative adversarial networks,” *arXiv preprint arXiv:1902.09599*.

MATTEI, P.-A., AND J. FRELLSEN (2019): “MIWAE: Deep generative modelling and imputation of incomplete data sets,” in *International conference on machine learning*, pp. 4413–4423. PMLR.

MORENO, P., P. HO, AND N. VASCONCELOS (2003): “A Kullback-Leibler divergence based kernel for SVM classification in multimedia applications,” *Advances in neural information processing systems*, 16.

MUZELLEC, B., J. JOSSE, C. BOYER, AND M. CUTURI (2020): “Missing data imputation using optimal transport,” in *International Conference on Machine Learning*, pp. 7130–7140. PMLR.

NAZABAL, A., P. M. OLMOS, Z. GHAHRAMANI, AND I. VALERA (2020): “Handling incomplete heterogeneous data using vaes,” *Pattern Recognition*, 107, 107501.

NGUYEN, X., M. J. WAINWRIGHT, AND M. I. JORDAN (2010): “Estimating divergence functionals and the likelihood ratio by convex risk minimization,” *IEEE Transactions on Information Theory*, 56(11), 5847–5861.

NOWOZIN, S., B. CSEKE, AND R. TOMIOKA (2016): “f-gan: Training generative neural samplers using variational divergence minimization,” *Advances in neural information processing systems*, 29.

PÉREZ-CRUZ, F. (2008): “Kullback-Leibler divergence estimation of continuous distributions,” in *2008 IEEE international symposium on information theory*, pp. 1666–1670. IEEE.

Schennach, S. M. (2020): "Mismeasured and unobserved variables," in *Handbook of Econometrics*, vol. 7, pp. 487–565. Elsevier.

Wang, Q., S. R. Kulkarni, and S. Verdú (2005): "Divergence estimation of continuous distributions based on data-dependent partitions," *IEEE Transactions on Information Theory*, 51(9), 3064–3074.

——— (2006): "A nearest-neighbor approach to estimating divergence between continuous random vectors," in *2006 IEEE International Symposium on Information Theory*, pp. 242–246. IEEE.

Woloszko, N. (2021): "Tracking GDP using Google Trends and machine learning: A new OECD model," *Central Banking*, 12, 12.

Yoon, J., J. Jordon, and M. Schaar (2018): "Gain: Missing data imputation using generative adversarial nets," in *International conference on machine learning*, pp. 5689–5698. PMLR.

Yoon, S., and S. Sull (2020): "GAMIN: Generative adversarial multiple imputation network for highly missing data," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 8456–8464.

A Appendix

A.1 Identification in distribution

This section presents the nonparametric identification results. We assume

Assumption 2 *There exists a random variable* X^* *with support* $\mathcal{X}^*$ *such that*

$$\begin{aligned} & f_{X^1,X^2,...,X^k,X^*} \\ = & f_{X^1|X^*}f_{X^2|X^*} \times ... \times f_{X^k|X^*}f_{X^*} \end{aligned}$$

We may consider the observables $(X^1, X^2, ..., X^k)$ as measurements of X^*. Here we use Hu and Schennach (2008) to show the uniqueness of $f(X^1, X^2, ..., X^k, X^*)$. We assume three of the k measurements are informative enough for the results in Hu and Schennach (2008). We assume

Assumption 3 *The joint distribution of* $(X^1, X^2, ..., X^k, X^*)$ *with* $k \geq 3$ *admits a bounded density with respect to the product measure of some dominating measure defined on their supports. All marginal and conditional densities are also bounded.*

Before introducing more assumptions, we define an integral operator corresponding to $f_{X^1|X^*}$, which maps f_{X^*} over support $\mathcal{X}^*$ to f_{X^1} over support $\mathcal{X}^1$. Suppose that we know both f_{X^*} and f_{X^1} are bounded and integrable. We define $\mathcal{L}^1_{bnd}(\mathcal{X}^*)$ as the set of bounded and integrable functions defined on $\mathcal{X}^*$, i.e.,

$$\begin{aligned} & \mathcal{L}^1_{bnd}(\mathcal{X}^*) \\ = & \left\{ g : \int_{\mathcal{X}^*} |g(x^*)|\, dx^* < \infty \text{ and } \sup_{x^* \in \mathcal{X}^*} |g(x^*)| < \infty \right\}. \end{aligned}$$

The linear operator can be defined as

$$\begin{aligned} L_{X^1|X^*} &: \mathcal{L}^1_{bnd}(\mathcal{X}^*) \to \mathcal{L}^1_{bnd}(\mathcal{X}^1) \\ \left(L_{X^1|X^*}h\right)(x) &= \int_{\mathcal{X}^*} f_{X^1|X^*}(x|x^*)h(x^*)dx^*. \end{aligned} \tag{8}$$

In order to identify the unknown distributions, we need the observables to be informative so that the following assumptions hold.

Assumption 4 *The operators* $L_{X^1|X^*}$ *and* $L_{X^2|X^1}$ *are injective.*[2]

Assumption 5 *For all* $\overline{x}^* \neq \widetilde{x}^*$ *in* $\mathcal{X}^*$*, the set* $\left\{x^3 : f_{X^3|X^*}(x^3|\overline{x}^*) \neq f_{X^3|X^*}(x^3|\widetilde{x}^*)\right\}$ *has positive probability.*

Assumption 6 *There exists a known functional* M *such that* $M\left[f_{X^1|X^*}(\cdot|x^*)\right] = x^*$ *for all* $x^* \in \mathcal{X}^*$*.*

The functional M may be the mean, mode, medium, or another quantile of the distribution $f_{X^1|X^*}(\cdot|x^*)$. The identification result may be summarized as follows:

Theorem 7 *Hu and Schennach (2008) Under assumptions 2, 3, 4, 5, and 6 in section A.1, the joint distribution* $f_{X^1,X^2,...,X^k}$ *uniquely determines the joint distribution* $f_{X^1,X^2,...,X^k,X^*}$*, which satisfies*

$$\begin{aligned} & f_{X^1,X^2,...,X^k,X^*} \\ = \; & f_{X^1|X^*} f_{X^2|X^*} \times ... \times f_{X^k|X^*} f_{X^*}. \end{aligned} \tag{9}$$

A.2 Convergence arguments

We present convergence arguments of our estimator here. Suppose our identification results suggest that our estimates $\hat{X}_i^*$ should have the same distribution (and variance) as X_i^*. Then the sample moments of $\hat{X}_i^*$ should converges to the true moments. In other words, we have

$$\frac{1}{N}\sum_{i=1}^{N}(\hat{X}_i^*)^2 - \frac{1}{N}\sum_{i=1}^{N}(X_i^*)^2 = o_p(1) \tag{10}$$

Such a condition implies the consistency of our estimator under following assumptions.

Theorem 8 *Suppose that the estimator* $\hat{X}_i^* = X_i^* + \delta_i$ *for* $i = 1, 2, ..., N$ *satisfies*

$$\frac{1}{N}\sum_{i=1}^{N} X_i^* \delta_i = o_p(1). \tag{11}$$

Then, the consistency of the sample moment in equation (10) implies that for any $\epsilon > 0$*, the sample proportion of large deviations goes to zero, i.e.,*

$$P_N\left(\left|\hat{X}_i^* - X_i^*\right| > \epsilon\right) := \frac{1}{N}\sum_{i=1}^{N} I(|\delta_i| > \epsilon) = o_p(1)$$

---

[2] $L_{X^2|X^1}$ is defined in the same way as $L_{X^1|X^*}$ in equation (8).

Proof: With $\hat{X}_i^* = X_i^* + \delta_i$, we have

$$\begin{aligned}
& \frac{1}{N}\sum_{i=1}^{N}(\hat{X}_i^*)^2 - \frac{1}{N}\sum_{i=1}^{N}(X_i^*)^2 \\
= \ & 2\frac{1}{N}\sum_{i=1}^{N} X_i^*\delta_i + \frac{1}{N}\sum_{i=1}^{N}(\delta_i)^2 \\
= \ & o_p(1) + \frac{1}{N}\sum_{i=1}^{N}(\delta_i)^2 \\
= \ & o_p(1)
\end{aligned}$$

Therefore,

$$\frac{1}{N}\sum_{i=1}^{N}(\delta_i)^2 = o_p(1)$$

Furthermore, we have

$$\frac{1}{N}\sum_{i=1}^{N}(\delta_i)^2 > \epsilon^2\frac{1}{N}\sum_{i=1}^{N} I(|\delta_i| > \epsilon)$$

Therefore, for any $\epsilon > 0$, we have

$$\frac{1}{N}\sum_{i=1}^{N} I(|\delta_i| > \epsilon) = o_p(1)$$

■

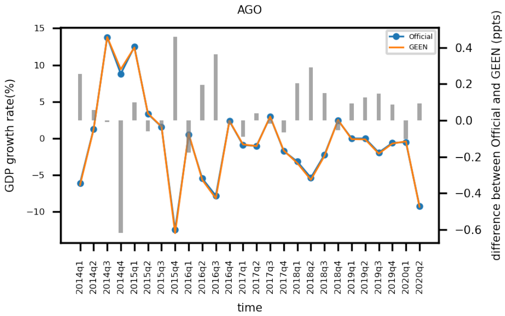
This result implies that the estimator for each observation is consistent w.r.t the sampling distribution. Theorem 2.4 in the main paper does not guarantee the consistency of $\hat{X}_i^*$ in a given observation, but the probability of a randomly-drew estimator $\hat{X}_i^*$ being consistent should converge to one.

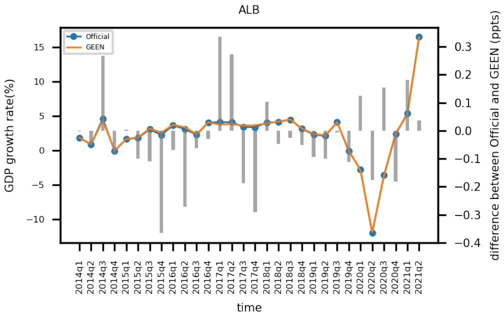
This result can be extended to more general deviations. For example, condition (11) may be replaced with

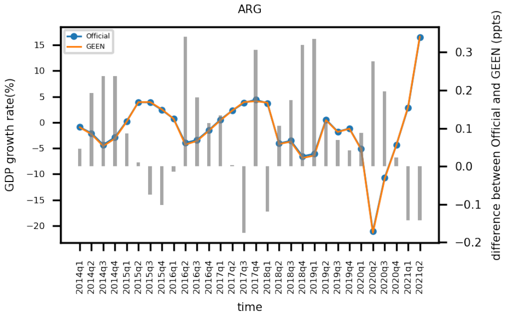
$$\frac{1}{N}\sum_{i=1}^{N} X_i^*\delta_i = c \times \frac{1}{N}\sum_{i=1}^{N}(\delta_i)^2 + o_p(1). \tag{12}$$

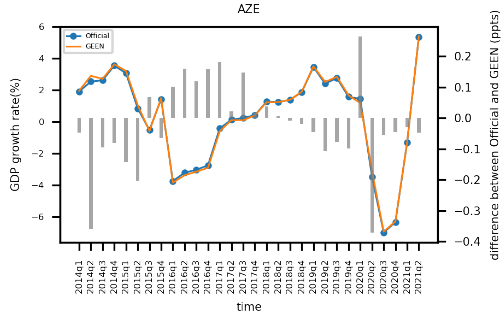
where c is a constant satisfying $c \neq -\frac{1}{2}$. That means the identification result remains in some cases where the deviations are correlated with the true values in the limit.

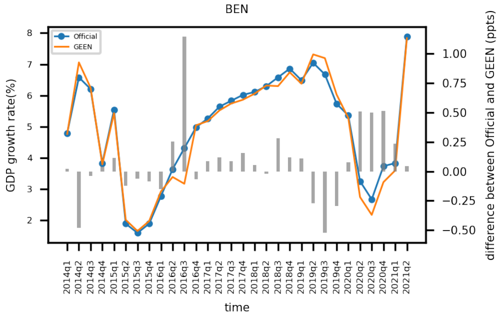
A.3 Results for GDP Refinement

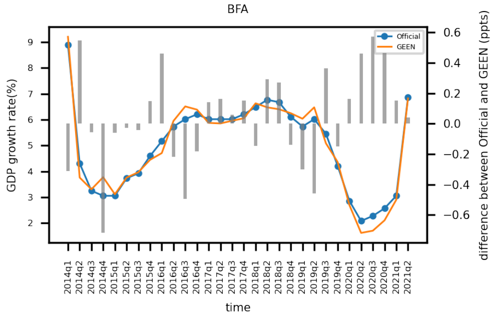

AGO
Official
GEEN
GDP growth rate(%)
difference between Official and GEEN (ppts)
time

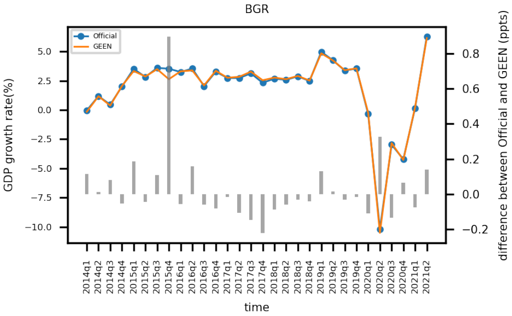

ALB
Official
GEEN
GDP growth rate(%)
difference between Official and GEEN (ppts)
time

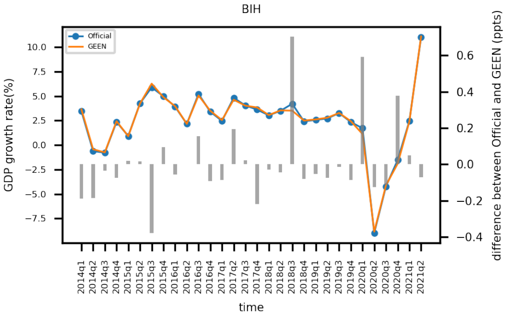

ARG
Official
GEEN
GDP growth rate(%)
difference between Official and GEEN (ppts)
time

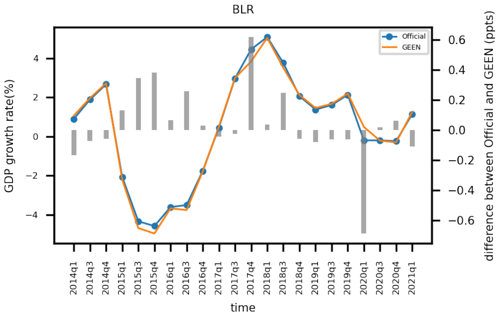

AZE
Official
GEEN
GDP growth rate(%)
difference between Official and GEEN (ppts)
time

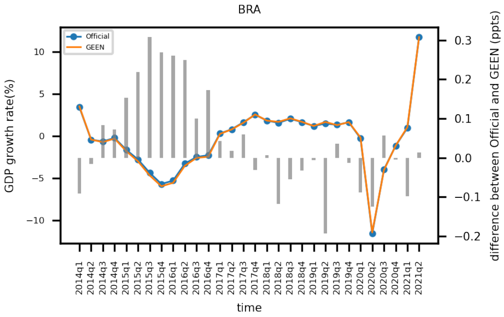

BEN
Official
GEEN
GDP growth rate(%)
difference between Official and GEEN (ppts)
time

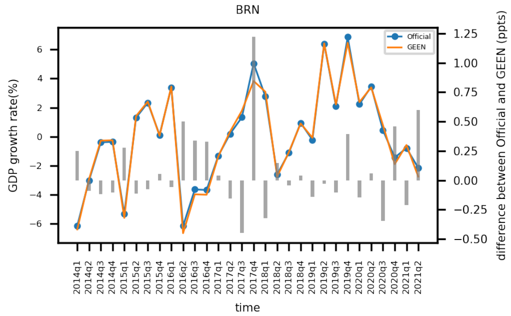

BFA
Official
GEEN
GDP growth rate(%)
difference between Official and GEEN (ppts)
time

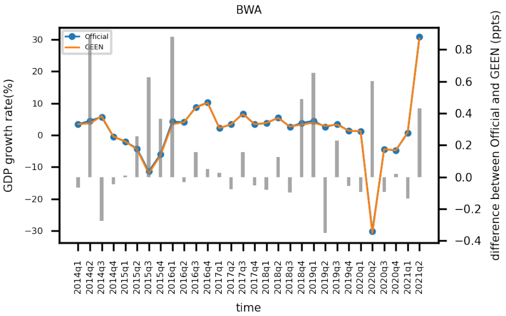

BGR
Official
GEEN
GDP growth rate(%)
difference between Official and GEEN (ppts)
time

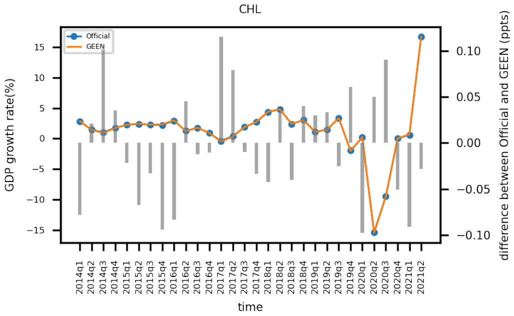

BIH
Official
GEEN
GDP growth rate(%)
difference between Official and GEEN (ppts)
time

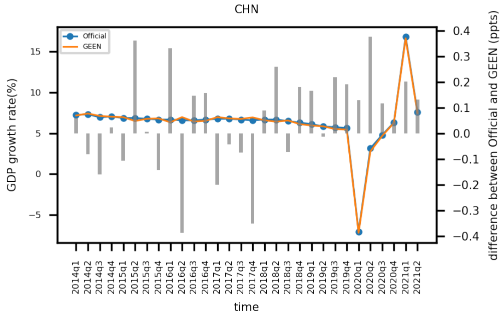

BLR
Official
GEEN
GDP growth rate(%)
difference between Official and GEEN (ppts)
time

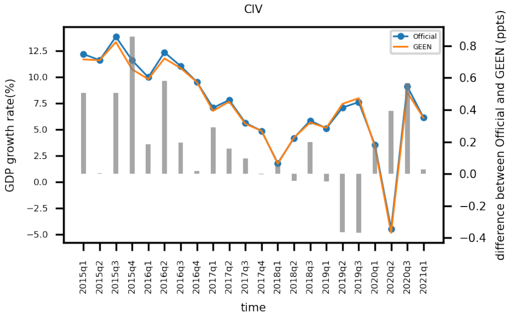

BRA
Official
GEEN
GDP growth rate(%)
difference between Official and GEEN (ppts)
time

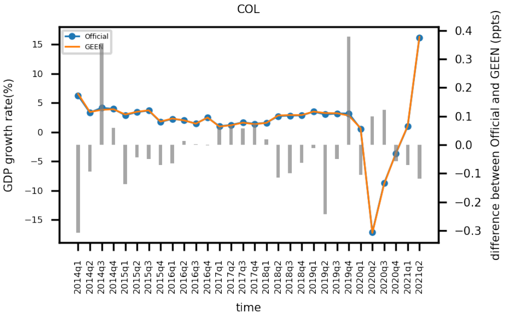

BRN
Official
GEEN
GDP growth rate(%)
6 4 2 0 −2 −4 −6
difference between Official and GEEN (ppts)
1.25 1.00 0.75 0.50 0.25 0.00 −0.25 −0.50
2014q1 2014q2 2014q3 2014q4 2015q1 2015q2 2015q3 2015q4 2016q1 2016q2 2016q3 2016q4 2017q1 2017q2 2017q3 2017q4 2018q1 2018q2 2018q3 2018q4 2019q1 2019q2 2019q3 2019q4 2020q1 2020q2 2020q3 2020q4 2021q1 2021q2
time

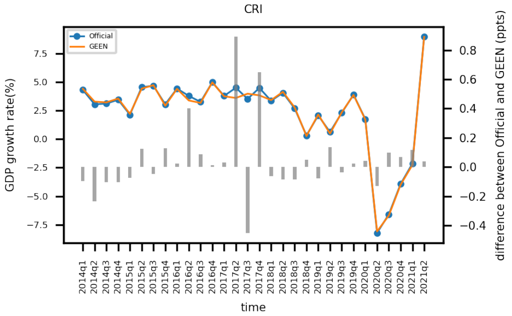

BWA
Official
GEEN
GDP growth rate(%)
30 20 10 0 −10 −20 −30
difference between Official and GEEN (ppts)
0.8 0.6 0.4 0.2 0.0 −0.2 −0.4
2014q1 2014q2 2014q3 2014q4 2015q1 2015q2 2015q3 2015q4 2016q1 2016q2 2016q3 2016q4 2017q1 2017q2 2017q3 2017q4 2018q1 2018q2 2018q3 2018q4 2019q1 2019q2 2019q3 2019q4 2020q1 2020q2 2020q3 2020q4 2021q1 2021q2
time

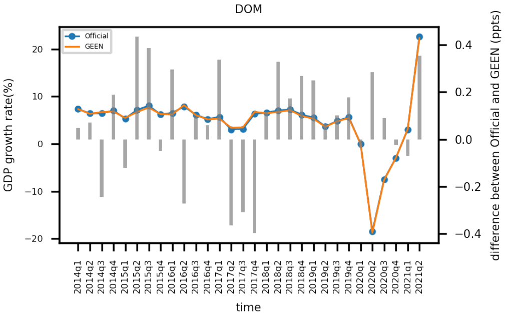

CHL
Official
GEEN
GDP growth rate(%)
15 10 5 0 −5 −10 −15
difference between Official and GEEN (ppts)
0.10 0.05 0.00 −0.05 −0.10
2014q1 2014q2 2014q3 2014q4 2015q1 2015q2 2015q3 2015q4 2016q1 2016q2 2016q3 2016q4 2017q1 2017q2 2017q3 2017q4 2018q1 2018q2 2018q3 2018q4 2019q1 2019q2 2019q3 2019q4 2020q1 2020q2 2020q3 2020q4 2021q1 2021q2
time

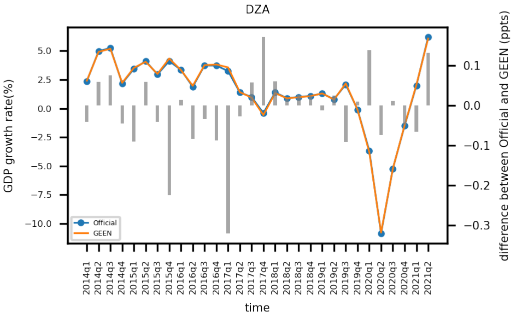

CHN
Official
GEEN
GDP growth rate(%)
15 10 5 0 −5
difference between Official and GEEN (ppts)
0.4 0.3 0.2 0.1 0.0 −0.1 −0.2 −0.3 −0.4
2014q1 2014q2 2014q3 2014q4 2015q1 2015q2 2015q3 2015q4 2016q1 2016q2 2016q3 2016q4 2017q1 2017q2 2017q3 2017q4 2018q1 2018q2 2018q3 2018q4 2019q1 2019q2 2019q3 2019q4 2020q1 2020q2 2020q3 2020q4 2021q1 2021q2
time

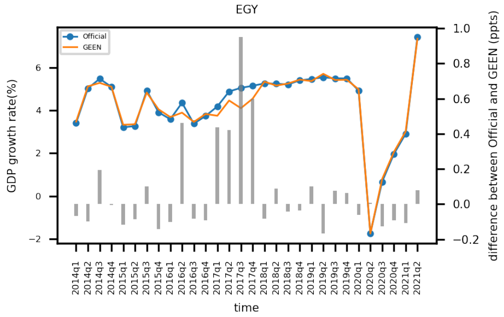

CIV
Official
GEEN
GDP growth rate(%)
12.5 10.0 7.5 5.0 2.5 0.0 −2.5 −5.0
difference between Official and GEEN (ppts)
0.8 0.6 0.4 0.2 0.0 −0.2 −0.4
2015q1 2015q2 2015q3 2015q4 2016q1 2016q2 2016q3 2016q4 2017q1 2017q2 2017q3 2017q4 2018q1 2018q2 2018q3 2018q4 2019q1 2019q2 2019q3 2019q4 2020q1 2020q2 2020q3 2021q1
time

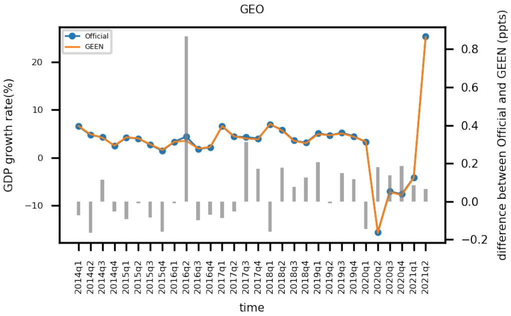

COL
Official
GEEN
GDP growth rate(%)
15 10 5 0 −5 −10 −15
difference between Official and GEEN (ppts)
0.4 0.3 0.2 0.1 0.0 −0.1 −0.2 −0.3
2014q1 2014q2 2014q3 2014q4 2015q1 2015q2 2015q3 2015q4 2016q1 2016q2 2016q3 2016q4 2017q1 2017q2 2017q3 2017q4 2018q1 2018q2 2018q3 2018q4 2019q1 2019q2 2019q3 2019q4 2020q1 2020q2 2020q3 2020q4 2021q1 2021q2
time

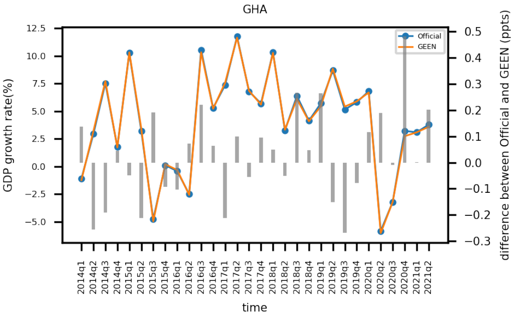

CRI
Official
GEEN
GDP growth rate(%)
7.5 5.0 2.5 0.0 −2.5 −5.0 −7.5
difference between Official and GEEN (ppts)
0.8 0.6 0.4 0.2 0.0 −0.2 −0.4
2014q1 2014q2 2014q3 2014q4 2015q1 2015q2 2015q3 2015q4 2016q1 2016q2 2016q3 2016q4 2017q1 2017q2 2017q3 2017q4 2018q1 2018q2 2018q3 2018q4 2019q1 2019q2 2019q3 2019q4 2020q1 2020q2 2020q3 2020q4 2021q1 2021q2
time

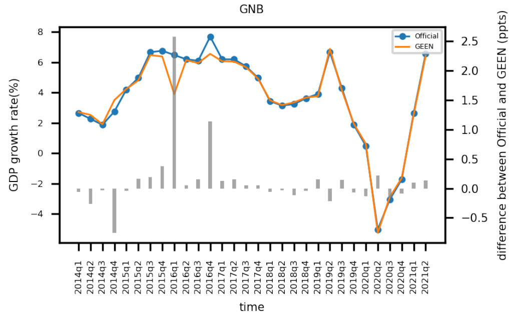

DOM
Official
GEEN
GDP growth rate(%)
20 10 0 −10 −20
difference between Official and GEEN (ppts)
0.4 0.2 0.0 −0.2 −0.4
2014q1 2014q2 2014q3 2014q4 2015q1 2015q2 2015q3 2015q4 2016q1 2016q2 2016q3 2016q4 2017q1 2017q2 2017q3 2017q4 2018q1 2018q2 2018q3 2018q4 2019q1 2019q2 2019q3 2019q4 2020q1 2020q2 2020q3 2020q4 2021q1 2021q2
time

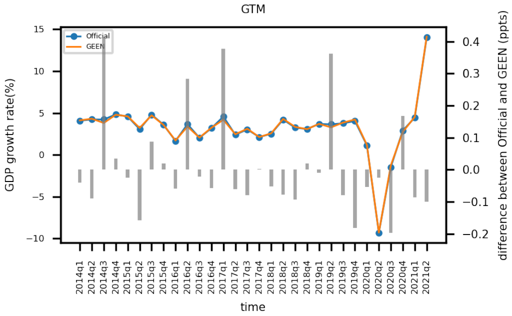

DZA
Official
GEEN
GDP growth rate(%)
5.0 2.5 0.0 −2.5 −5.0 −7.5 −10.0
difference between Official and GEEN (ppts)
0.1 0.0 −0.1 −0.2 −0.3
2014q1 2014q2 2014q3 2014q4 2015q1 2015q2 2015q3 2015q4 2016q1 2016q2 2016q3 2016q4 2017q1 2017q2 2017q3 2017q4 2018q1 2018q2 2018q3 2018q4 2019q1 2019q2 2019q3 2019q4 2020q1 2020q2 2020q3 2020q4 2021q1 2021q2
time

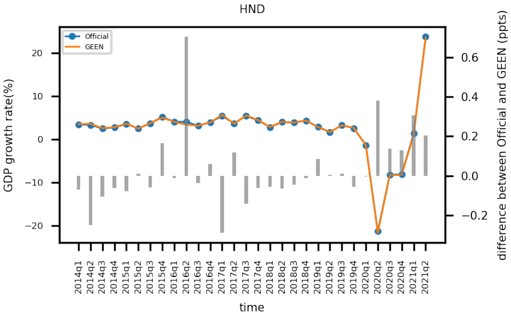

EGY
Official
GEEN
GDP growth rate(%)
6 4 2 0 −2
difference between Official and GEEN (ppts)
1.0 0.8 0.6 0.4 0.2 0.0 −0.2
2014q1 2014q2 2014q3 2014q4 2015q1 2015q2 2015q3 2015q4 2016q1 2016q2 2016q3 2016q4 2017q1 2017q2 2017q3 2017q4 2018q1 2018q2 2018q3 2018q4 2019q1 2019q2 2019q3 2019q4 2020q1 2020q2 2020q3 2020q4 2021q1 2021q2
time

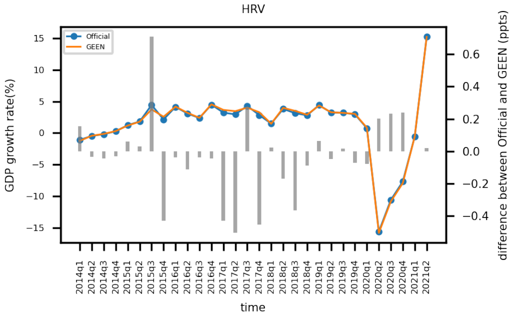

GEO
Official
GEEN
GDP growth rate(%)
20
10
0
−10
difference between Official and GEEN (ppts)
0.8
0.6
0.4
0.2
0.0
−0.2
2014q1 2014q2 2014q3 2014q4 2015q1 2015q2 2015q3 2015q4 2016q1 2016q2 2016q3 2016q4 2017q1 2017q2 2017q3 2017q4 2018q1 2018q2 2018q3 2018q4 2019q1 2019q2 2019q3 2019q4 2020q1 2020q2 2020q3 2020q4 2021q1 2021q2
time

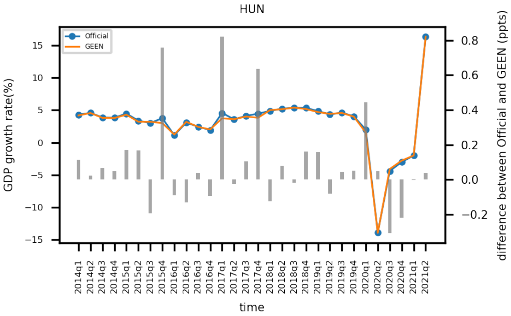

GHA
Official
GEEN
GDP growth rate(%)
12.5
10.0
7.5
5.0
2.5
0.0
−2.5
−5.0
difference between Official and GEEN (ppts)
0.5
0.4
0.3
0.2
0.1
0.0
−0.1
−0.2
−0.3
2014q1 2014q2 2014q3 2014q4 2015q1 2015q2 2015q3 2015q4 2016q1 2016q2 2016q3 2016q4 2017q1 2017q2 2017q3 2017q4 2018q1 2018q2 2018q3 2018q4 2019q1 2019q2 2019q3 2019q4 2020q1 2020q2 2020q3 2020q4 2021q1 2021q2
time

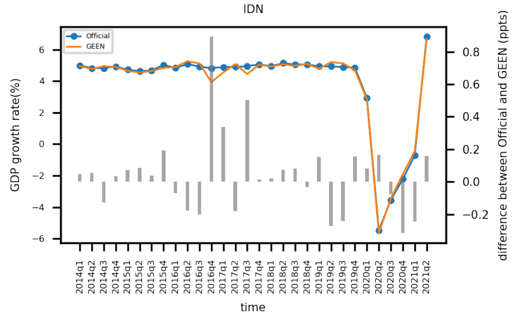

GNB
Official
GEEN
GDP growth rate(%)
8
6
4
2
0
−2
−4
difference between Official and GEEN (ppts)
2.5
2.0
1.5
1.0
0.5
0.0
−0.5
2014q1 2014q2 2014q3 2014q4 2015q1 2015q2 2015q3 2015q4 2016q1 2016q2 2016q3 2016q4 2017q1 2017q2 2017q3 2017q4 2018q1 2018q2 2018q3 2018q4 2019q1 2019q2 2019q3 2019q4 2020q1 2020q2 2020q3 2020q4 2021q1 2021q2
time

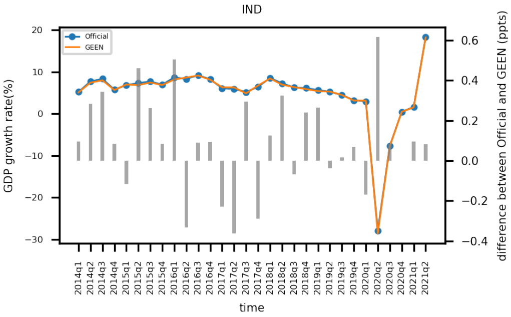

GTM
Official
GEEN
GDP growth rate(%)
15
10
5
0
−5
−10
difference between Official and GEEN (ppts)
0.4
0.3
0.2
0.1
0.0
−0.1
−0.2
2014q1 2014q2 2014q3 2014q4 2015q1 2015q2 2015q3 2015q4 2016q1 2016q2 2016q3 2016q4 2017q1 2017q2 2017q3 2017q4 2018q1 2018q2 2018q3 2018q4 2019q1 2019q2 2019q3 2019q4 2020q1 2020q2 2020q3 2020q4 2021q1 2021q2
time

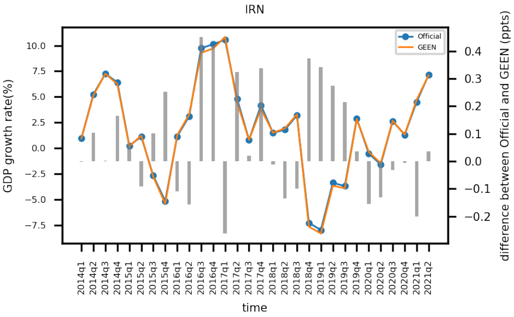

HND
Official
GEEN
GDP growth rate(%)
20
10
0
−10
−20
difference between Official and GEEN (ppts)
0.6
0.4
0.2
0.0
−0.2
2014q1 2014q2 2014q3 2014q4 2015q1 2015q2 2015q3 2015q4 2016q1 2016q2 2016q3 2016q4 2017q1 2017q2 2017q3 2017q4 2018q1 2018q2 2018q3 2018q4 2019q1 2019q2 2019q3 2019q4 2020q1 2020q2 2020q3 2020q4 2021q1 2021q2
time

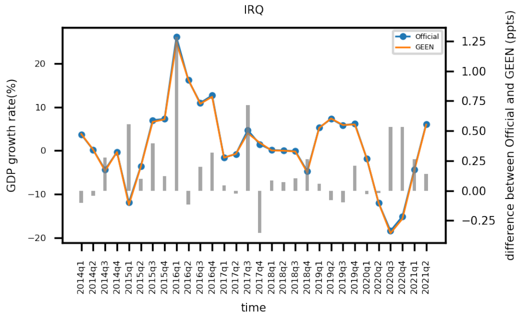

HRV
Official
GEEN
GDP growth rate(%)
15
10
5
0
−5
−10
−15
difference between Official and GEEN (ppts)
0.6
0.4
0.2
0.0
−0.2
−0.4
2014q1 2014q2 2014q3 2014q4 2015q1 2015q2 2015q3 2015q4 2016q1 2016q2 2016q3 2016q4 2017q1 2017q2 2017q3 2017q4 2018q1 2018q2 2018q3 2018q4 2019q1 2019q2 2019q3 2019q4 2020q1 2020q2 2020q3 2020q4 2021q1 2021q2
time

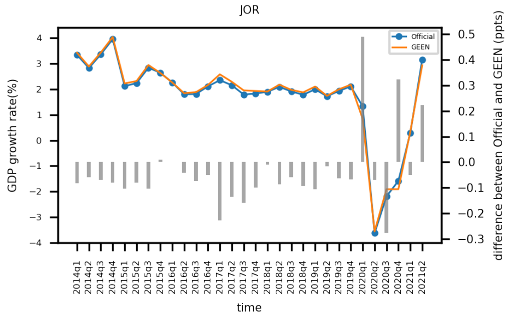

HUN
Official
GEEN
GDP growth rate(%)
15
10
5
0
−5
−10
−15
difference between Official and GEEN (ppts)
0.8
0.6
0.4
0.2
0.0
−0.2
2014q1 2014q2 2014q3 2014q4 2015q1 2015q2 2015q3 2015q4 2016q1 2016q2 2016q3 2016q4 2017q1 2017q2 2017q3 2017q4 2018q1 2018q2 2018q3 2018q4 2019q1 2019q2 2019q3 2019q4 2020q1 2020q2 2020q3 2020q4 2021q1 2021q2
time

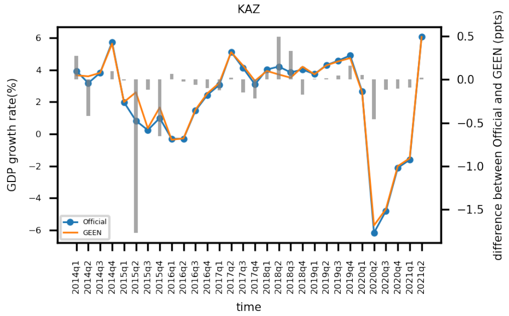

IDN
Official
GEEN
GDP growth rate(%)
6
4
2
0
−2
−4
−6
difference between Official and GEEN (ppts)
0.8
0.6
0.4
0.2
0.0
−0.2
2014q1 2014q2 2014q3 2014q4 2015q1 2015q2 2015q3 2015q4 2016q1 2016q2 2016q3 2016q4 2017q1 2017q2 2017q3 2017q4 2018q1 2018q2 2018q3 2018q4 2019q1 2019q2 2019q3 2019q4 2020q1 2020q2 2020q3 2020q4 2021q1 2021q2
time

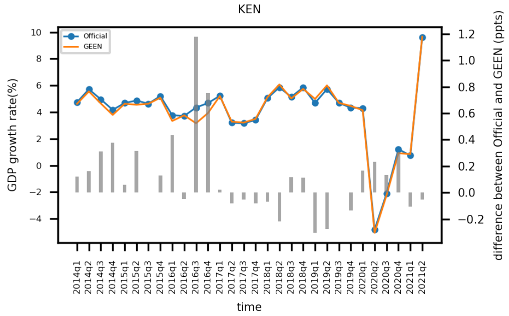

IND
Official
GEEN
GDP growth rate(%)
20
10
0
−10
−20
−30
difference between Official and GEEN (ppts)
0.6
0.4
0.2
0.0
−0.2
−0.4
2014q1 2014q2 2014q3 2014q4 2015q1 2015q2 2015q3 2015q4 2016q1 2016q2 2016q3 2016q4 2017q1 2017q2 2017q3 2017q4 2018q1 2018q2 2018q3 2018q4 2019q1 2019q2 2019q3 2019q4 2020q1 2020q2 2020q3 2020q4 2021q1 2021q2
time

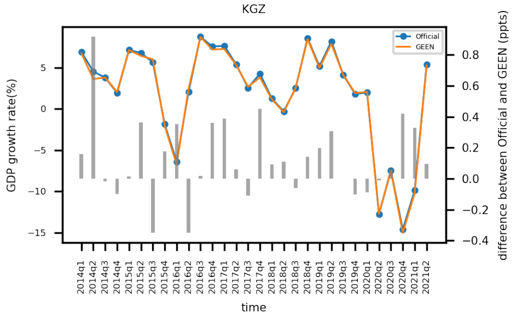

IRN
Official
GEEN
GDP growth rate(%)
10.0
7.5
5.0
2.5
0.0
−2.5
−5.0
−7.5
difference between Official and GEEN (ppts)
0.4
0.3
0.2
0.1
0.0
−0.1
−0.2
2014q1 2014q2 2014q3 2014q4 2015q1 2015q2 2015q3 2015q4 2016q1 2016q2 2016q3 2016q4 2017q1 2017q2 2017q3 2017q4 2018q1 2018q2 2018q3 2018q4 2019q1 2019q2 2019q3 2019q4 2020q1 2020q2 2020q3 2020q4 2021q1 2021q2
time

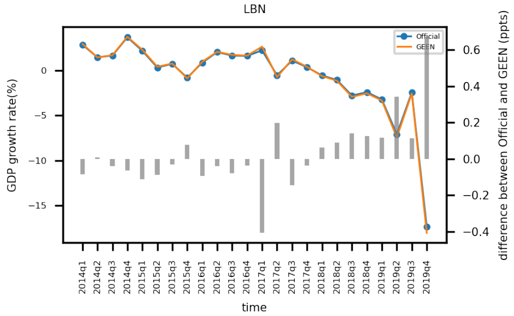

IRQ
Official
GEEN
GDP growth rate(%)
difference between Official and GEEN (ppts)
time
20
10
0
−10
−20
1.25
1.00
0.75
0.50
0.25
0.00
−0.25
2014q1 2014q2 2014q3 2014q4 2015q1 2015q2 2015q3 2015q4 2016q1 2016q2 2016q3 2016q4 2017q1 2017q2 2017q3 2017q4 2018q1 2018q2 2018q3 2018q4 2019q1 2019q2 2019q3 2019q4 2020q1 2020q2 2020q3 2020q4 2021q1 2021q2

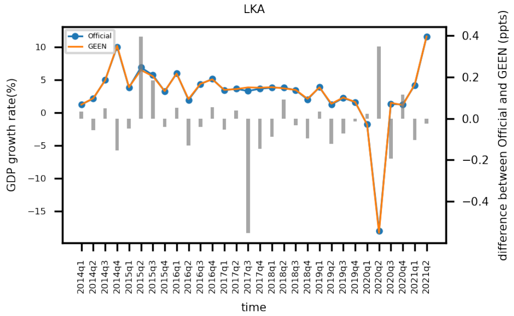

JOR
Official
GEEN
GDP growth rate(%)
difference between Official and GEEN (ppts)
time
4
3
2
1
0
−1
−2
−3
−4
0.5
0.4
0.3
0.2
0.1
0.0
−0.1
−0.2
−0.3
2014q1 2014q2 2014q3 2014q4 2015q1 2015q2 2015q3 2015q4 2016q1 2016q2 2016q3 2016q4 2017q1 2017q2 2017q3 2017q4 2018q1 2018q2 2018q3 2018q4 2019q1 2019q2 2019q3 2019q4 2020q1 2020q2 2020q3 2020q4 2021q1 2021q2

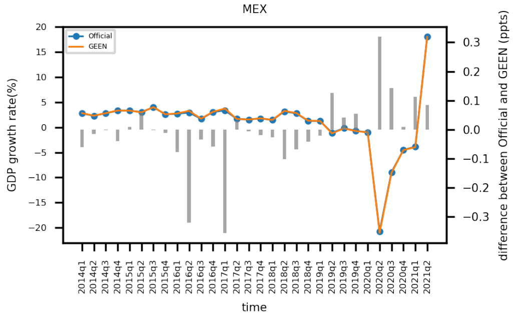

KAZ
Official
GEEN
GDP growth rate(%)
difference between Official and GEEN (ppts)
time
6
4
2
0
−2
−4
−6
0.5
0.0
−0.5
−1.0
−1.5
2014q1 2014q2 2014q3 2014q4 2015q1 2015q2 2015q3 2015q4 2016q1 2016q2 2016q3 2016q4 2017q1 2017q2 2017q3 2017q4 2018q1 2018q2 2018q3 2018q4 2019q1 2019q2 2019q3 2019q4 2020q1 2020q2 2020q3 2020q4 2021q1 2021q2

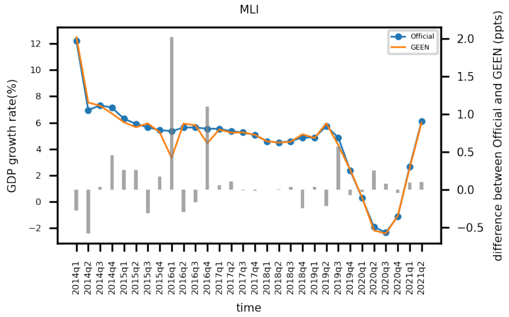

KEN
Official
GEEN
GDP growth rate(%)
difference between Official and GEEN (ppts)
time
10
8
6
4
2
0
−2
−4
1.2
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
2014q1 2014q2 2014q3 2014q4 2015q1 2015q2 2015q3 2015q4 2016q1 2016q2 2016q3 2016q4 2017q1 2017q2 2017q3 2017q4 2018q1 2018q2 2018q3 2018q4 2019q1 2019q2 2019q3 2019q4 2020q1 2020q2 2020q3 2020q4 2021q1 2021q2

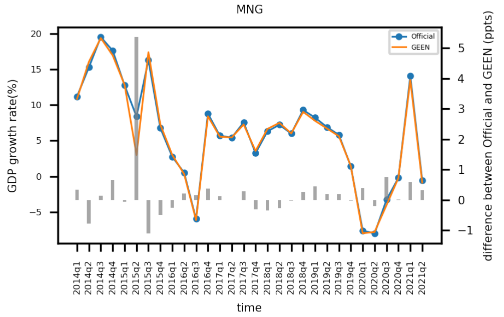

KGZ
Official
GEEN
GDP growth rate(%)
difference between Official and GEEN (ppts)
time
5
0
−5
−10
−15
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4
2014q1 2014q2 2014q3 2014q4 2015q1 2015q2 2015q3 2015q4 2016q1 2016q2 2016q3 2016q4 2017q1 2017q2 2017q3 2017q4 2018q1 2018q2 2018q3 2018q4 2019q1 2019q2 2019q3 2019q4 2020q1 2020q2 2020q3 2020q4 2021q1 2021q2

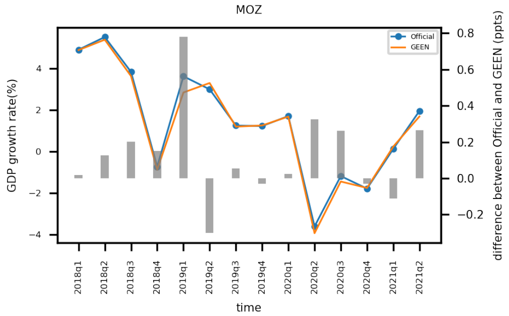

LBN
Official
GEEN
GDP growth rate(%)
difference between Official and GEEN (ppts)
time
0
−5
−10
−15
0.6
0.4
0.2
0.0
−0.2
−0.4
2014q1 2014q2 2014q3 2014q4 2015q1 2015q2 2015q3 2015q4 2016q1 2016q2 2016q3 2016q4 2017q1 2017q2 2017q3 2017q4 2018q1 2018q2 2018q3 2018q4 2019q1 2019q2 2019q3 2019q4

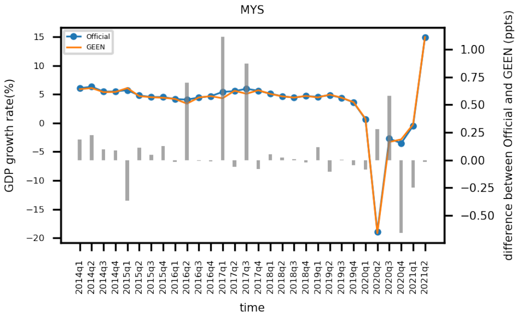

LKA
Official
GEEN
GDP growth rate(%)
difference between Official and GEEN (ppts)
time
10
5
0
−5
−10
−15
0.4
0.2
0.0
−0.2
−0.4
2014q1 2014q2 2014q3 2014q4 2015q1 2015q2 2015q3 2015q4 2016q1 2016q2 2016q3 2016q4 2017q1 2017q2 2017q3 2017q4 2018q1 2018q2 2018q3 2018q4 2019q1 2019q2 2019q3 2019q4 2020q1 2020q2 2020q3 2020q4 2021q1 2021q2

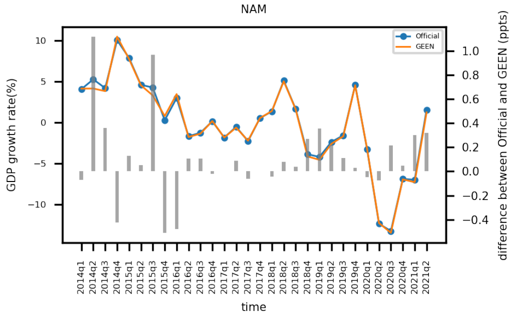

MEX
Official
GEEN
GDP growth rate(%)
difference between Official and GEEN (ppts)
time
20
15
10
5
0
−5
−10
−15
−20
0.3
0.2
0.1
0.0
−0.1
−0.2
−0.3
2014q1 2014q2 2014q3 2014q4 2015q1 2015q2 2015q3 2015q4 2016q1 2016q2 2016q3 2016q4 2017q1 2017q2 2017q3 2017q4 2018q1 2018q2 2018q3 2018q4 2019q1 2019q2 2019q3 2019q4 2020q1 2020q2 2020q3 2020q4 2021q1 2021q2

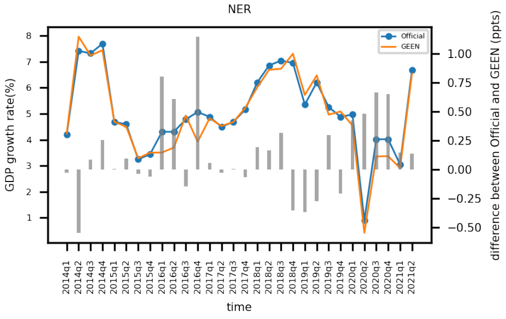

MLI
Official
GEEN
GDP growth rate(%)
difference between Official and GEEN (ppts)
time
12
10
8
6
4
2
0
−2
2.0
1.5
1.0
0.5
0.0
−0.5
2014q1 2014q2 2014q3 2014q4 2015q1 2015q2 2015q3 2015q4 2016q1 2016q2 2016q3 2016q4 2017q1 2017q2 2017q3 2017q4 2018q1 2018q2 2018q3 2018q4 2019q1 2019q2 2019q3 2019q4 2020q1 2020q2 2020q3 2020q4 2021q1 2021q2

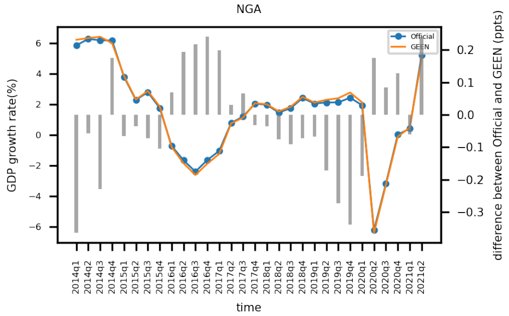

MNG
Official
GEEN
GDP growth rate(%)
difference between Official and GEEN (ppts)
time
20
15
10
5
0
−5
5
4
3
2
1
0
−1
2014q1 2014q2 2014q3 2014q4 2015q1 2015q2 2015q3 2015q4 2016q1 2016q2 2016q3 2016q4 2017q1 2017q2 2017q3 2017q4 2018q1 2018q2 2018q3 2018q4 2019q1 2019q2 2019q3 2019q4 2020q1 2020q2 2020q3 2020q4 2021q1 2021q2

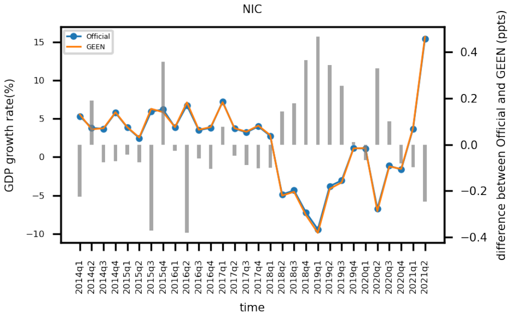

MOZ
Official
GEEN
GDP growth rate(%)
−4 −2 0 2 4
difference between Official and GEEN (ppts)
−0.2 0.0 0.2 0.4 0.6 0.8
2018q1 2018q2 2018q3 2018q4 2019q1 2019q2 2019q3 2019q4 2020q1 2020q2 2020q3 2020q4 2021q1 2021q2
time

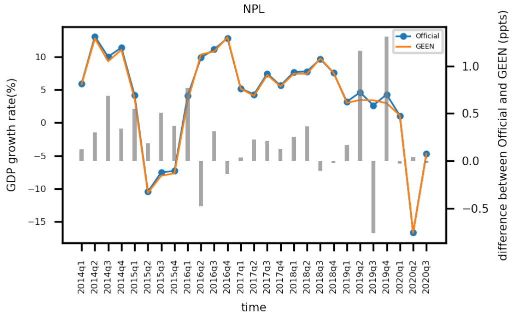

MYS
Official
GEEN
GDP growth rate(%)
−20 −15 −10 −5 0 5 10 15
difference between Official and GEEN (ppts)
−0.50 −0.25 0.00 0.25 0.50 0.75 1.00
2014q1 … 2021q2
time

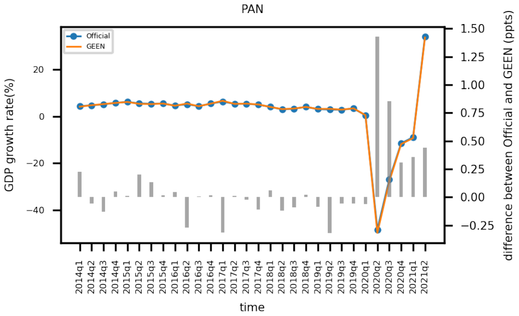

NAM
Official
GEEN
GDP growth rate(%)
−10 −5 0 5 10
difference between Official and GEEN (ppts)
−0.4 −0.2 0.0 0.2 0.4 0.6 0.8 1.0
2014q1 … 2021q2
time

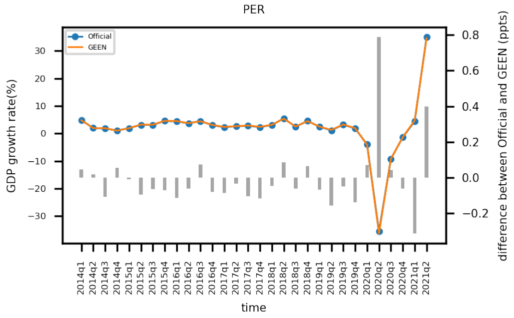

NER
Official
GEEN
GDP growth rate(%)
1 2 3 4 5 6 7 8
difference between Official and GEEN (ppts)
−0.50 −0.25 0.00 0.25 0.50 0.75 1.00
2014q1 … 2021q2
time

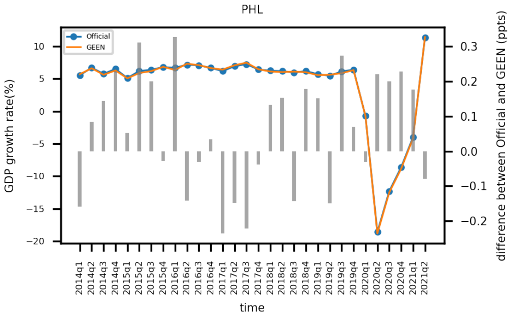

NGA
Official
GEEN
GDP growth rate(%)
−6 −4 −2 0 2 4 6
difference between Official and GEEN (ppts)
−0.3 −0.2 −0.1 0.0 0.1 0.2
2014q1 … 2021q2
time

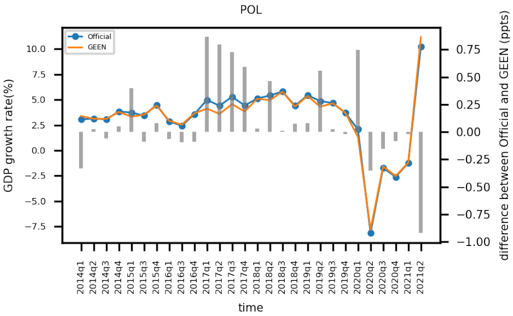

NIC
Official
GEEN
GDP growth rate(%)
−10 −5 0 5 10 15
difference between Official and GEEN (ppts)
−0.4 −0.2 0.0 0.2 0.4
2014q1 … 2021q2
time

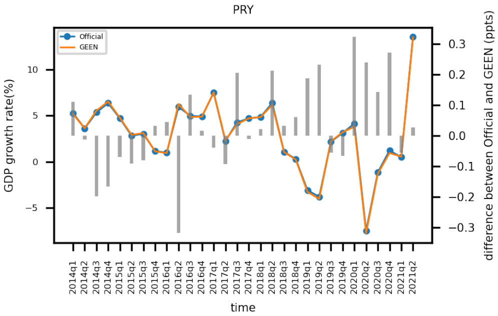

NPL
Official
GEEN
GDP growth rate(%)
−15 −10 −5 0 5 10
difference between Official and GEEN (ppts)
−0.5 0.0 0.5 1.0
2014q1 … 2020q3
time

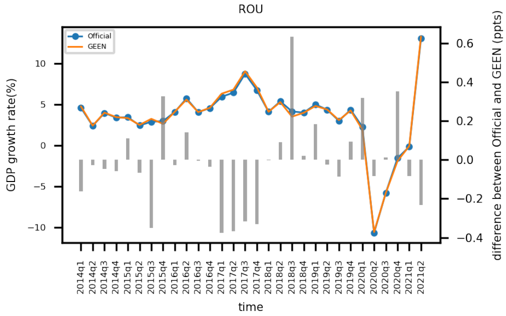

PAN
Official
GEEN
GDP growth rate(%)
−40 −20 0 20
difference between Official and GEEN (ppts)
−0.25 0.00 0.25 0.50 0.75 1.00 1.25 1.50
2014q1 … 2021q2
time

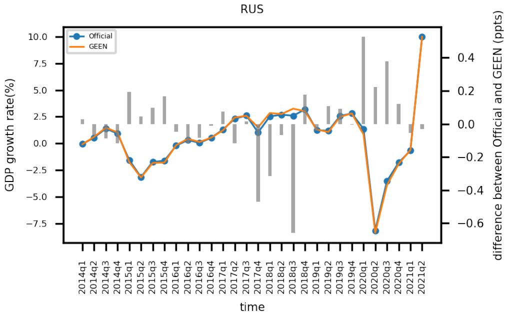

PER
Official
GEEN
GDP growth rate(%)
−30 −20 −10 0 10 20 30
difference between Official and GEEN (ppts)
−0.2 0.0 0.2 0.4 0.6 0.8
2014q1 … 2021q2
time

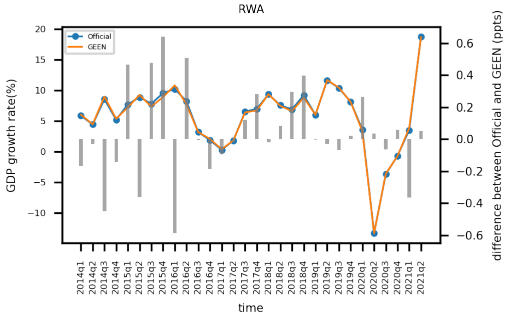

PHL
Official
GEEN
GDP growth rate(%)
−20 −15 −10 −5 0 5 10
difference between Official and GEEN (ppts)
−0.2 −0.1 0.0 0.1 0.2 0.3
2014q1 … 2021q2
time

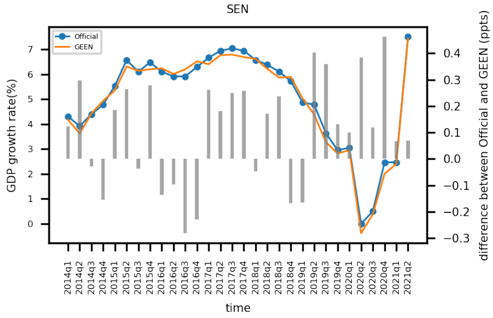

POL
Official
GEEN
GDP growth rate(%)
difference between Official and GEEN (ppts)
time
10.0 7.5 5.0 2.5 0.0 −2.5 −5.0 −7.5
0.75 0.50 0.25 0.00 −0.25 −0.50 −0.75 −1.00
2014q1 2014q2 2014q3 2014q4 2015q1 2015q2 2015q3 2015q4 2016q1 2016q2 2016q3 2016q4 2017q1 2017q2 2017q3 2017q4 2018q1 2018q2 2018q3 2018q4 2019q1 2019q2 2019q3 2019q4 2020q1 2020q2 2020q3 2020q4 2021q1 2021q2

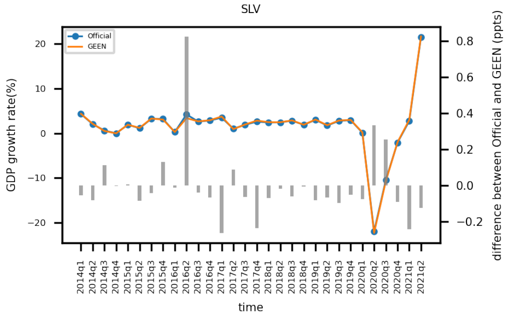

PRY
Official
GEEN
GDP growth rate(%)
difference between Official and GEEN (ppts)
time
10 5 0 −5
0.3 0.2 0.1 0.0 −0.1 −0.2 −0.3
2014q1 2014q2 2014q3 2014q4 2015q1 2015q2 2015q3 2015q4 2016q1 2016q2 2016q3 2016q4 2017q1 2017q2 2017q3 2017q4 2018q1 2018q2 2018q3 2018q4 2019q1 2019q2 2019q3 2019q4 2020q1 2020q2 2020q3 2020q4 2021q1 2021q2

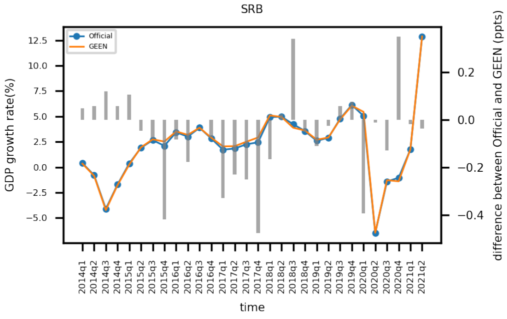

ROU
Official
GEEN
GDP growth rate(%)
difference between Official and GEEN (ppts)
time
10 5 0 −5 −10
0.6 0.4 0.2 0.0 −0.2 −0.4
2014q1 2014q2 2014q3 2014q4 2015q1 2015q2 2015q3 2015q4 2016q1 2016q2 2016q3 2016q4 2017q1 2017q2 2017q3 2017q4 2018q1 2018q2 2018q3 2018q4 2019q1 2019q2 2019q3 2019q4 2020q1 2020q2 2020q3 2020q4 2021q1 2021q2

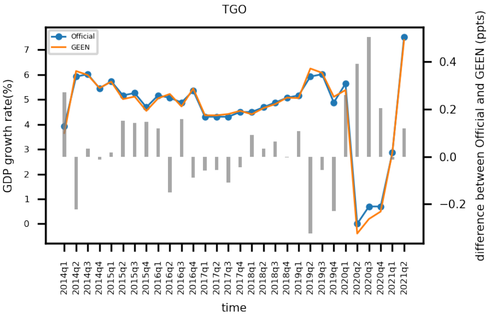

RUS
Official
GEEN
GDP growth rate(%)
difference between Official and GEEN (ppts)
time
10.0 7.5 5.0 2.5 0.0 −2.5 −5.0 −7.5
0.4 0.2 0.0 −0.2 −0.4 −0.6
2014q1 2014q2 2014q3 2014q4 2015q1 2015q2 2015q3 2015q4 2016q1 2016q2 2016q3 2016q4 2017q1 2017q2 2017q3 2017q4 2018q1 2018q2 2018q3 2018q4 2019q1 2019q2 2019q3 2019q4 2020q1 2020q2 2020q3 2020q4 2021q1 2021q2

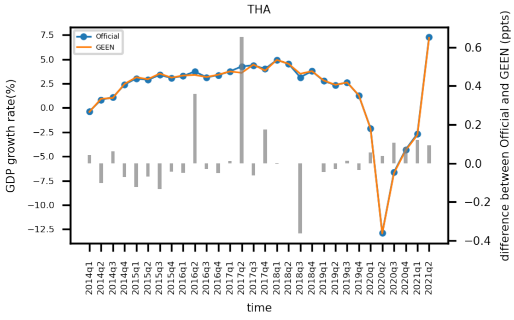

RWA
Official
GEEN
GDP growth rate(%)
difference between Official and GEEN (ppts)
time
20 15 10 5 0 −5 −10
0.6 0.4 0.2 0.0 −0.2 −0.4 −0.6
2014q1 2014q2 2014q3 2014q4 2015q1 2015q2 2015q3 2015q4 2016q1 2016q2 2016q3 2016q4 2017q1 2017q2 2017q3 2017q4 2018q1 2018q2 2018q3 2018q4 2019q1 2019q2 2019q3 2019q4 2020q1 2020q2 2020q3 2020q4 2021q1 2021q2


SEN
Official
GEEN
GDP growth rate(%)
difference between Official and GEEN (ppts)
time
7 6 5 4 3 2 1 0
0.4 0.3 0.2 0.1 0.0 −0.1 −0.2 −0.3
2014q1 2014q2 2014q3 2014q4 2015q1 2015q2 2015q3 2015q4 2016q1 2016q2 2016q3 2016q4 2017q1 2017q2 2017q3 2017q4 2018q1 2018q2 2018q3 2018q4 2019q1 2019q2 2019q3 2019q4 2020q1 2020q2 2020q3 2020q4 2021q1 2021q2


SLV
Official
GEEN
GDP growth rate(%)
difference between Official and GEEN (ppts)
time
20 10 0 −10 −20
0.8 0.6 0.4 0.2 0.0 −0.2
2014q1 2014q2 2014q3 2014q4 2015q1 2015q2 2015q3 2015q4 2016q1 2016q2 2016q3 2016q4 2017q1 2017q2 2017q3 2017q4 2018q1 2018q2 2018q3 2018q4 2019q1 2019q2 2019q3 2019q4 2020q1 2020q2 2020q3 2020q4 2021q1 2021q2


SRB
Official
GEEN
GDP growth rate(%)
difference between Official and GEEN (ppts)
time
12.5 10.0 7.5 5.0 2.5 0.0 −2.5 −5.0
0.2 0.0 −0.2 −0.4
2014q1 2014q2 2014q3 2014q4 2015q1 2015q2 2015q3 2015q4 2016q1 2016q2 2016q3 2016q4 2017q1 2017q2 2017q3 2017q4 2018q1 2018q2 2018q3 2018q4 2019q1 2019q2 2019q3 2019q4 2020q1 2020q2 2020q3 2020q4 2021q1 2021q2


TGO
Official
GEEN
GDP growth rate(%)
difference between Official and GEEN (ppts)
time
7 6 5 4 3 2 1 0
0.4 0.2 0.0 −0.2
2014q1 2014q2 2014q3 2014q4 2015q1 2015q2 2015q3 2015q4 2016q1 2016q2 2016q3 2016q4 2017q1 2017q2 2017q3 2017q4 2018q1 2018q2 2018q3 2018q4 2019q1 2019q2 2019q3 2019q4 2020q1 2020q2 2020q3 2020q4 2021q1 2021q2


THA
Official
GEEN
GDP growth rate(%)
difference between Official and GEEN (ppts)
time
7.5 5.0 2.5 0.0 −2.5 −5.0 −7.5 −10.0 −12.5
0.6 0.4 0.2 0.0 −0.2 −0.4
2014q1 2014q2 2014q3 2014q4 2015q1 2015q2 2015q3 2015q4 2016q1 2016q2 2016q3 2016q4 2017q1 2017q2 2017q3 2017q4 2018q1 2018q2 2018q3 2018q4 2019q1 2019q2 2019q3 2019q4 2020q1 2020q2 2020q3 2020q4 2021q1 2021q2

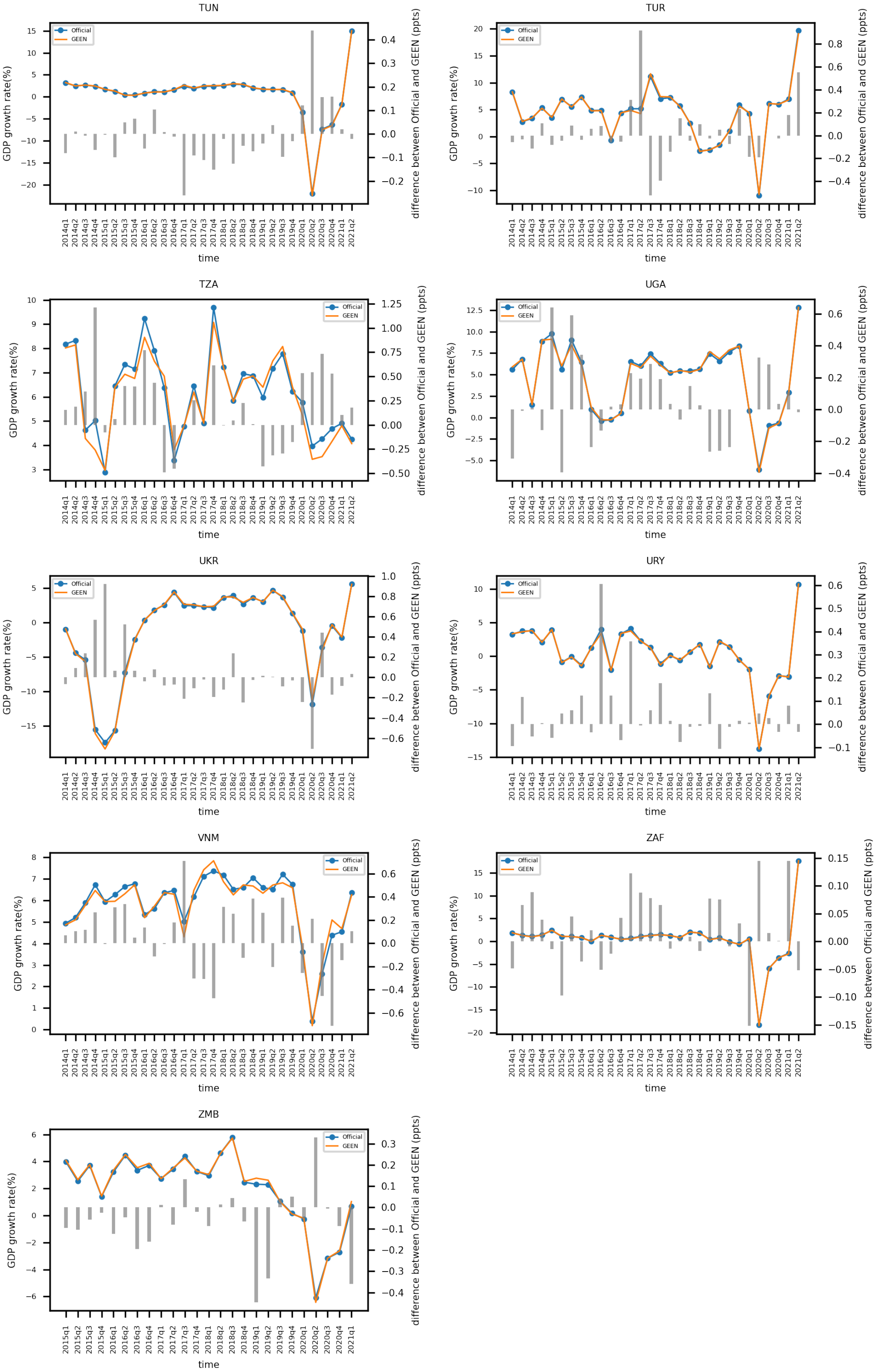


Figure 9: Official GDP and Generated Underlying GDP