

Greenhouse Emission Forecast as an Inverse Stochastic Problem: A Cross-Entropy Econometrics Approach

S. BWANAKARE

University of Information Technology and Management in Rzeszow, Sucharskiego 2, 35-225 Rzeszow, Poland

This paper proposes the non-extensive entropy econometric approach to predict regional cross-industry greenhouse emissions within a country, based on imperfect knowledge of industrial and regional aggregates. The solution of this stochastic inverse problem is applied to Poland. Non-extensive entropy should remain a valuable device for econometric modelling even in the case of low frequency series since outputs provided by the Gibbs–Shannon entropy approach correspond to the Tsallis entropy limiting case of the Gaussian law when the Tsallis q -parameter equals unity. We, therefore, set up a q -Tsallis–Kullback–Leibler entropy criterion function with a priori consistency constraints, including the environmental Kuznets econometric model and regular conditions. As in the case of Shannon–Gibbs-based entropy models, we found that the Tsallis entropy estimator also belongs to the family of Stein estimators, meaning that smaller probabilities are shrunk and higher probabilities dominate in the solution space. Fortunately, adding more pertinent data to the model priors will enhance parameter precision and then allow for the recovery of the real influence of smaller events. The q -Tsallis–Kullback–Leibler entropy index is computed for different scenarios of the Kuznets model. The model outputs continue to conform to empirical expectations. In spite of the close to unity q -Tsallis parameter, this Tsallis related approach reflects higher stability for parameter computation in comparison with the Shannon–Gibbs entropy econometrics technique.

DOI: [10.12693/APhysPolA.127.A-13](https://doi.org/10.12693/APhysPolA.127.A-13)

PACS: 89.65.Gh, 89.70.cf

1. Introduction

In many real world situations, theoreticians and empiricists observe two or more quantifiable multivariate stochastic systems and would like to infer unknown cross-correlation between their elements. To illustrate such an ill-posed stochastic inverse problem, this paper proposes q -Tsallis information divergence-related approach to forecast a matrix of industrial greenhouse emission within a country by region based on defective or incomplete regional or industrial aggregates. In particular, as often occurs in empirical investigations, the proposed model analyses the situation where row and column totals do not match as a consequence of, for instance, systematic measurement error and/or contradictory information from different sources.

Studies in various scientific fields have successfully applied power law (PL) formalism in high frequency series modelling (e.g., [1, 2]). Interestingly enough, PL — of which Gaussian law constitutes a converging case for Tsallis- q parameter equal to unity — can be extended to low frequency series (e.g., [3]), illustrated in the case study below by one period two dimension in Table I.

For decades, statistical and mathematical tools for dealing with ill-posed inverse problem system recovery have been sought in diverse fields — model parameter estimation, medical imaging, modelling in the life sciences, oil and mineral deposit exploration, shape optimization, and so on. The basic* model for treating ill-posed inverse

problems consists in solving an integral equation of the first kind. We formulate it — in the context of the model to be later developed — as follows:

$$G(\zeta) = \int_D f(\beta)h(\beta, \zeta)d\beta + b(\zeta), \quad (1)$$

where G means the observed row or column totals, f designates the unknown regional cross-industry matrix of greenhouse pollutants, D defines the Hilbert support space of the model, h is the transformation kernel linking measures G and f , b explains random errors.

Literature on different methodologies dedicated to ill-posed inverse problem recovery is expansive. Besides the well-known Tikhonov regularization theory [4], Gibbs–Shannon–Jaynes [5, 6] maximum (minimum) entropy principle and its recent extensions [7, 8] remain the most used techniques for solving this class of problem. The general principal which applies to both approaches is the association of a linear or nonlinear least square problem with a regularization term (*a priori* or additional information) so as to obtain a *well-posed* problem. Additionally, the Gibbs–Shannon–Jaynes maximum (minimum) entropy principle formalism searches for *global* regularity — related to the second law of thermodynamics — while yielding the smoothest reconstructions consistent with the available data in the Bayesian spirit. It is worth adding that to ensure the possibility of an optimal solution, a discrete form of Eq. (1) has been retained for empirical illustration.

e-mail: swanakare@wsiz.rzeszow.pl

*For instance, limiting ourselves to usual signal or imagery

cases, the basic equation (Eq. 1) can be extended to include, e.g., measurement system impulse response.

As far as the social sciences are concerned, a number of other techniques have been tried for this class of inverse problem. One can cite the Moore–Penrose pseudo-inverse problem approach with theoretical limitations for a certain class of ill-behaved matrix or the bi-proportional RAS approach and its extension [9]. While this last technique requires an initial transaction matrix, it offers a poorer solution when the investigated model is *stochastic*. Golan, et al. [8] have shown lower performance of the Markov chain model with respect to generalized Gibbs–Shannon entropy for this class of inverse problem. Next, the Bayesian approach has shown its relative superiority, particularly one connected with the maximum entropy principle. A model of the neuronal class could also be suggested. However, it is not based on a compact theory, its application is time consuming, and its outputs are not always guaranteed. It should be emphasized that the Shannon–Gibbs relative entropy model has been successfully applied for updating and balancing social accounting matrices [9]. However, on theoretical grounds, this assumes that entropy is a positive function of the number of possible states, extensive, and it then neglects the possibility of inter-correlations among the states and their impact. In a recent paper, authors [10] have proven the convergence of two standard regularization techniques toward two special Tsallis q -values. For $q = 2$, one should obtain the Tikhonov regularization and for $q = 1$, one gets the classical Boltzmann–Gibbs–Shannon formulation for entropy. The central point is that a PL displays, besides its well-known scaling law, a set of interesting characterizations related to its aggregative properties, in that it is conserved under addition, multiplication, polynomial transformation, minimum, and maximum [11].

As we deal with an ill-conditioned inverse problem, we need all three regularity conditions (existence, uniqueness, and stability of the solution) to be simultaneously fulfilled. As will be shown below, obtaining a satisfactory, unique and robust solution requires constrained consistency information [12, 13] — as a Bayes efficient information processing rule — besides which the relative non-extensive entropy criterion function will be defined.

While, generally, the *existence* and *uniqueness* conditions remain reachable through regular a priori restrictions, *stability* of optimal solution, corrupted by random or systematic errors, is much harder to retrieve. In essence, the problem is finding, among an infinity of distributions satisfying all the imposed constraints, the one that best replicates the data generating system (DGS). As far as maximum entropy formalism is concerned, by virtue of Jaynes contribution [5], reasonably, the best candidate should be the one that most reduces uncertainty about the system. Likewise, by virtue of Kullback-Leibler information divergence metrics [8, 14], the best candidate should be those posteriors which satisfy all constraining conditions while diverging least from the priors.

We will introduce into the model the standard Kuznet’s environmental economic function [15–17], and

its relevance with respect to solution stability — over a given time frame — will be examined.

2. The model and confidence interval area

We apply the usual[†] discrete form of the q -Generalized Kullback-Leibler relative entropy (Eq. 5). Thanks to the new entropy constraining data, the model updates initial information and provides new outputs. We need to reparameterize the generalized linear model (Eq. 2), which plays the role of restrictions. Transaction elements to be forecasted can be meaningfully presented by columns as ratios explaining each region’s average cross-industry weight of the greenhouse emission. Thus, each coefficient can be treated as a probability and the coefficient total by column must sum up to unity. In this instance, the parameterization support space coincides with the probability space. In such a case, the accuracy of estimated parameters is higher as there is non-loss of information from this a priori data [18]. In any event, let us succinctly present the general procedure of reparameterization in the case of a general linear inverse model:

$$Y = X'\beta + \varepsilon, \quad (2)$$

where unknown β parameter values are not necessarily constrained between 0 and 1, which suggests the necessity for reparameterization. The term ε is an unobservable disturbance term, plausibly with finite variance, owing to the nature of the phenomena involved, exhibiting observation errors from empirical measurement or from random shocks possibly driven by PL. The variable Y is the observed data — with errors — from the unknown DGS of greenhouse pollutant per industry[‡], and X may stand for the known — with uncertainty — regional emitted greenhouse pollutants through the relational parameter matrix β and the unobservable disturbance ε to be estimated through the observable error components e . Unlike the classical econometric models, no constraining hypotheses — about, e.g., error distribution — is needed. In particular, as we are concerned with an ill-behaved inverse problem, the number of parameters to be estimated should be higher than the observed data points, and the quality of collected information data should be low. The process of the true system recovery requires the entropy objective function to incorporate all of the constraining consistency moments. Thus, referring to the properties of the relative entropy principle, each new piece of constraining information will reduce the entropy level of the system in accordance with the degree of data consistency with the system. For this two dimensional space inverse problem, among an unlimited number of model solution candidates, the best solution will result

[†]The generalized Bregman Kullback-Leibler may be the alternative version of this model.

[‡]We note that this variable could also represent the regional greenhouse pollutant system. However, the reason for such a choice will be explained later while commenting on outputs.

from identifying the one that — in terms of probability — best simulates the data generating system.

Taking each β_{kl} ($k = 1 \dots K$, $l = 1 \dots L$) as a discrete, random variable with compact support [8] and $2 < M < \infty$ possible outcomes, one can estimate it by B_{kl} , that is:

$$B_{kl} = \sum_{m=1}^M p_{klm} v_{klm}, \quad (3)$$

where p_{klm} is the probability of outcome v_{klm} , and the probabilities must be non-negative and sum up to one. Similarly, by treating each element $e_{\bullet l}$ (affecting industry sector total uncertainty) of e as a finite and discrete random variable with compact support and $2 < J < \infty$ possible outcomes centred around zero, we can express $e_{\bullet l}$ as:

$$e_{\bullet l} = \sum_{j=1 \dots J} r_{\bullet lj} z_{\bullet lj}, \quad (4)$$

As mentioned, one can assume that each row total prior has been assessed with errors $\omega_{k\bullet}$ and then a similar support space must be built as follows:

$$\omega_{k\bullet} = \sum_{s=1}^S \mu_{k\bullet s} v_{k\bullet s},$$

where $r_{\bullet lj}$ and $\mu_{k\bullet s}$ are respectively the probabilities of outcomes z_{klj} and v_{ls} on the support spaces $j = 1 \dots J$ and $s = 1 \dots S$. Thus, indexes $k = 1 \dots K$ and $l = 1 \dots L$ represent, respectively, the number of row and column totals, data of which are supposed to have been estimated with errors. Furthermore, the error terms e and ω have been empirically fixed around the empirical standard error of the explained variables and stand for the a priori Bayesian hypothesis. The choice of error bounds naturally depends on their own proprieties. In this study, their sets have been specified through Chebychev's inequality [19] with support space bounds varying between -3 and $+3$. It should be remembered that posterior probabilities within the support space may display a non-Gaussian distribution class.

The element v_{klm} constitutes an a priori information provided by the researcher while p_{klm} is an unknown probability generating the true parameter β_{kl} the value of which must be determined by solving a relative non-extensive entropy problem. In matrix notation, let us rewrite $\beta = VP$ with $p_{klm} \geq 0$ and $\sum_{k=1}^K \sum_{j>2 \dots J} p_{klm} = 1$. Also, let $e = rz$ with $r_{\bullet lj} \geq 0$ and $\sum_{j>2 \dots J} r_{\bullet lj} = 1$ for K and L the number of rows and columns and J the number of data points over the support space for the error terms inside the regional cross-industry matrix. The same normality conditions can be easily formulated for the random vector of column totals. Then, the Tsallis cross-entropy econometric estimator can be stated as:

$$\begin{aligned} \min H_q(p||p^0, r||r^0, \mu||\mu^0) \equiv \\ \alpha \sum p_{klm} \frac{[p_{klm}/p_{klm}^0]^{q-1} - 1}{q-1} \end{aligned}$$

$$\begin{aligned} + \beta \sum r_{\bullet lj} \frac{[r_{\bullet lj}/r_{\bullet lj}^0]^{q-1} - 1}{q-1} + \dots \\ + \delta \sum \mu_{k\bullet s} \frac{[\mu_{k\bullet s}/\mu_{k\bullet s}^0]^{q-1} - 1}{q-1}. \end{aligned} \quad (5)$$

Subject to

$$Y_{\bullet l} = \sum_k ([Y_{\bullet l} P_{kl}]' + e_{\bullet l}) = \left[\sum_k \left[\left(\sum_{m>2}^M Y_{\bullet l} v_{klm} (p_{klm}^q) \right)' + \sum_{j=1 \dots J} r_{\bullet lj} z_{\bullet lj} \right] \right], \quad (6)$$

$$H_{k\bullet} = C_{k\bullet} (X_{k\bullet} + \omega_{k\bullet}) =$$

$$C_{k\bullet} \left(X_{k\bullet} + \sum_{s=1}^S \mu_{k\bullet s} v_{k\bullet s} \right), \quad (7)$$

$$\sum_{k \dots K} H_{k\bullet} = \sum_{l \dots L} Y_{\bullet l}, \quad (8)$$

$$\sum_{k=1}^K \sum_{j>2 \dots J} p_{klj} = 1, \quad (9)$$

$$\sum_{j>2 \dots J} r_{\bullet lj} = 1, \quad (10)$$

$$\sum_{s>2 \dots S} \mu_{k\bullet s} = 1,$$

where Y_l means each total by column l (observed values by industry l , including unknown error), H_k means each total by row (observed values by region k) after having adjusted by C_k , X_k — total pollutant emission by region out of unknown error, p_{kl} — probabilistic structure of greenhouse emissions across industry and region, C_k — a random scaling factor approximated, in this model, by the Kuznet coefficient, \bullet means variable related on the total row or column, according to context.

Non-extensive statistics has employed a number of constraining forms in which expectations may be defined. The above model applies the Curado-Tsallis (C-T) constraints [13, 20], the general form of which is as follows:

$$\langle y_q \rangle = \sum_i p_i^q y_i.$$

The real q , as previously stated, stands for the Tsallis parameter. As suggested by Table I, the coefficient $C_{k\bullet}$ remains a random scaling factor owing to non-consistent information displayed by the prior row and column totals known with uncertainty. Thus, besides random errors proposed in Eq. (4), we have introduced in the above model a scaling random factor $C_{k\bullet}$ in Eq. (7), so that sums of row and column totals could balance. The rationale for such a scaling random coefficient follows. In many situations, statistical data on different items are not only collected from different sources but explained in different units. In the present study, as commented

on below, both data from column totals and row totals are explained in the same units but at different scale and with a different degree of random statistical error. This leads to an impossibility to balance, as in theory is expected, overall totals of rows (Y_l) and columns (X_k). Thus, a random factor $C_{k\bullet}$ allows for such an error-correct term, resulting in Eq. (8). As it will be clear when we present outputs from model estimation, this error-correct term seems to be correlated with the Kuznets coefficient explained below. The t index value refers to the time period during which Kuznets's coefficient has been computed. According to Kuznets's environmental econometric theory, the quantity of emitted pollutants is a nonlinear function of national or regional economic welfare. In the first phases of development, the emitted quantity is smaller, and it steadily grows as a function of industrialisation. However, at a certain level of welfare, the emitted quantity per unit of produced GDP starts to decline. In the present model, the $C_{k\bullet}$ values concern sixteen Polish provinces and have been computed for the time period 1998–2005 (see Table II). Thus, in the above model [Eqs. (5)–(10)], their values will play the role of Bayesian new data, discriminating in favour of new inferential evidence. Since the greenhouse emission model has been built for the year 2005, we expect that the $C_{k\bullet}$ prior values of 2005 should have much more impact on the model outputs than those from 2004 or earlier. We should remember that when no new data are incorporated into the cross-entropy model, outputs correspond to those of maximum entropy formulation, with no additional prior, either, except normality conditions.

Above, $H_q(p||p^0, r||r^0, \mu||\mu^0)$ is nonlinear and measures the entropy in the model. Relative entropies of the three independent terms (the three posteriors p , r and μ and the corresponding priors p^0 , r^0 and μ^0 , respectively) are then summed up, eventually using the weights α , β , δ . These are real positives summing up to unity under the given restrictions. The first term of “precision” accounts for deviations of the estimated parameters from the prior (generally defined under a support space). The second and the third terms of “prediction ex post” account for the empirical error term as a difference between predicted and observed data values of the model.

We need to find the minimum divergence between the priors and the posteriors while the imposed restrictions are fulfilled. As will be the case in the application below, the first component of the criterion function may concern the parameter structure of the table, the second component errors on row totals, and the last component may concern errors on column totals.

Note that the model estimates and their variances should be affected not only by the support length but by its space scaling effect, i.e., the number of affected point values [8]. The higher the number of these points, the better the prior information about the system.

Next, the random errors ω_i (see Eq. (7)) explains errors in data gathering and processing not necessarily featuring

Gaussian distribution. As mentioned, the error prior initialization may have a significant impact on the posterior. Widening the error bound or decreasing observed points over the support space will result in posterior distribution shrinkage towards the prior values. This happens because widening support space is equivalent to less discriminating information. Algebraically, as proven in [8], widening the error bound will decrease the optimal Lagrange multiplier on the constrained model. As a result, the impact of a multiplier obtained in this way on the criterion function will be reduced. Traditionally, as far as Bayesian and relative entropy formulations are concerned, it may be worth mentioning that both models will lead to similar results when and only when the true expected error related to the data generating system is equal to zero within a support space symmetric around zero. Similarly, Gibbs–Shannon cross-entropy and Tsallis non-extensive cross-entropy outputs will correspond when errors embodied in the model are not correlated and then evolve within the Gaussian basin of attraction [e.g., 21–23]. Furthermore, the magnitudes of parameter or error bounds will affect the level of estimated parameter shrinkage towards the prior distribution. This level is measurable through the Lagrange multiplier obtained from the constraining model solution (Eqs. (5)–(10)). In the limit, when the multiplier tends toward zero, posterior values of the parameter will converge towards its prior distribution, explaining no additional information in the consistency constraints.

As far as parameter confidence interval area is concerned Eq. (11) reflects the non additivity property of Tsallis entropy for two — plausibly — independent systems, one related to the parameter probability distribution and the other to the error disturbance probability distribution:

$$S(\hat{P}r) = [S(\hat{p} + \hat{r}) = \{[S(\hat{p}) + S(\hat{r})] + (1 - q) S(\hat{p}) S(\hat{r})\}, \quad (11)$$

where:

$$S(\hat{p}) = - \left[1 - \sum_k \sum_m (p_{klm})^q \right] / [K (M^{1-q} - 1)]$$

and

$$S(\hat{r}) = - \left[\left(1 - \sum_l \sum_j r^q \bullet l_j \right) \right] / [L (J^{1-q} - 1)].$$

$S(\hat{P}r)$ is then the sum of normalized entropy related to parameters of the model $S(\hat{p})$ and to disturbance term $S(\hat{r})$. Likewise, the latter value $S(\hat{r})$ is derived for all observations l , with J the number of data points on the support space of estimated probabilities r related to the error term.

The values of these normalized entropy indexes vary between zero and one. Its values, near to one, indicate a poor informative variable while lower values are an indication of better informative estimated parameter $\hat{\beta}_{kl}$ about the model.

3. Recovery of Polish regional greenhouse emissions by industry

We will illustrate the empirical foundation of the theoretical model developed in the above sections through an application to the case of Poland. Table I — which the cross row and column symbols correspond to Eq. (6) — is presented to illustrate to what extent the researcher may dispose of information limited in quantity and quality before thus solving an inverse, ill-behaved problem, like the one presented in this article.

According to Eurostat, “The use of a common unit — kilogramme of CO₂ equivalent — allows the relative effect of different gases to be compared and combined. The quantity of emissions of individual greenhouse gases are converted into CO₂ equivalents and then aggregated. A single kilogramme of methane has 21 times the global warming effect of a kilogram of carbon dioxide, and a kilogram of nitrous oxide has 310 times the global warming effect of carbon dioxide”.

TABLE I

Polish industrial and regional greenhouse emission (thousands of tons CO₂) for the period 2005.

Polish regions	Industry and services							Total tons CO ₂
	Energy industry	Transport industry	Transport	Processing industry	Agriculture	Waste	Other	
Lódzkie	$Y_{\bullet 1}P_{11}$	$Y_{\bullet 2}P_{12}$	$Y_{\bullet l}P_{1l}$..	$Y_{\bullet L}P_{1L}$	37525.965
Mazowieckie	$Y_{\bullet 1}P_{21}$	$Y_{\bullet l}P_{2l}$..	$Y_{\bullet L}P_{2L}$	27229.675
Małopolskie	13787.91
Śląskie	40830.217
Lubelskie	5283.674
Podkarpackie	3519.443
Podlaskie	1799.787
Świętokrzyskie	9148.159
Lubuskie	$Y_{\bullet 1}P_{k1}$	$Y_{\bullet l}P_{kl}$..	$Y_{\bullet L}P_{kL}$	2189.278
Wielkopolskie	16987.325
Zachodniopomorskie	8071.549
Dolnośląskie	18048.75
Opolskie	13272.882
Kujawsko-pomorskie	8145.494
Pomorskie	6365.964
Warmińsko-mazurskie	$Y_{\bullet 1}P_{K1}$	$Y_{\bullet l}P_{Kl}$..	$Y_{\bullet L}P_{KL}$	1500.113
Total tons CO ₂	180369	32469	36443	28877	34581	9437	67787	

Sources: Eurostat and Polish Statistical Office (GUS).

We have at our disposal two pieces of aggregated information about greenhouse emissions from each Polish industry sector and from each province. In this problem, resulting from Table I [24, 25], we dispose of 112 degrees of freedom related to the number of cross-table greenhouse pollutant quantities to be estimated, and then, excluding one row and one column, the Kuznets coefficient vector $C_{k\bullet}$, treated in the model as a random variable, is included. If we were to deal with the k row and l column totals, this would then suggest a direct extension of the Jaynes dice problem [5] from one to two dimension discrete space, leading to $(k - 1) \times (l - 1)$ degree of freedom, which would illustrate the case of the standard inverse problem. Unfortunately, in the present case, row and column totals come from two different statistical sources displaying contradictory information. To exemplify this fact, let us observe (Table I) the discrepancy

between row total and column total sums which amount to 176 257, suggesting a difference of over 82%. These row total data (last row) come from the Eurostat Institute, and the column total data (last column) from GUS, the Polish national statistical institute. In each case, information concerns the period 2005. This kind of ill-conditioned inverse problem best fits the philosophy naturally embodied in the maximum entropy principle. The above total discrepancy probably results from two distinct sources. The first is different random measurement error from different institutions, and the second could be systematic measurement error, owing to different measurement hypotheses (scale error) applied by each institution.

The model outputs computed with the GAMS code are displayed in Table II. The prior probability p^{oklm} in Eq. (5) has been computed thanks to information from

TABLE II

Greenhouse pollutant post-entropy outputs (in thousands).

	Energy industry	Transport industry	Transport	Processing industry	Agricultural industry	Waste	Others	Total tons CO ₂
Lódzkie	29990	5701	6399	5071	6072	1657	11903	66793
Mazowieckie	32558	6204	6963	5518	6608	1803	12951	72605
Małopolskie	21987	4137	4643	3679	4406	1202	8637	48692
Śląskie	14025	2581	2897	2295	2749	750	5388	30685
Lubelskie	14850	2742	3078	2439	2921	797	5725	32552
Podkarpackie	11137	2017	2263	1793	2148	586	4210	24153
Podlaskie	11538	2095	2351	1863	2231	609	4373	25061
Świętokrzyskie	7932	1390	1560	1236	1480	404	2902	16903
Lubuskie	7094	1226	1376	1091	1306	356	2560	15009
Wielkopolskie	7153	1238	1389	1101	1318	360	2584	15142
Zachodniopomorskie	821	967	1086	860	1030	281	2019	7063
Dolnośląskie	3747	802	901	713	854	233	1676	8926
Opolskie	821	539	604	480	574	159	1120	4294
Kujawsko-pomorskie	821	333	373	296	354	97	694	2966
Pomorskie	821	273	307	243	291	79	571	2583
Warmińsko-mazurskie	821	228	256	203	243	66	476	2293
Total tons CO ₂	166113	32472	36445	28880	34584	9438	67788	

Sources: own computations.

TABLE III

Prior coefficients of greenhouse pollutants tons per one million unit of GDP over the period 1998–2005.

	1998	1999	2000	2001	2002	2003	2004	2005
Lódzkie	1.460	1.183	0.927	1.049	1.056	1.040	0.715	0.697
Mazowieckie	0.228	0.192	0.175	0.166	0.163	0.169	0.155	0.147
Małopolskie	0.346	0.305	0.276	0.257	0.245	0.241	0.219	0.218
Śląskie	0.490	0.431	0.386	0.367	0.391	0.394	0.371	0.355
Lubelskie	0.250	0.214	0.171	0.151	0.180	0.170	0.162	0.156
Podkarpackie	0.175	0.154	0.130	0.112	0.097	0.126	0.116	0.107
Podlaskie	0.151	0.145	0.108	0.112	0.107	0.111	0.097	0.089
Świętokrzyskie	0.862	0.761	0.699	0.622	0.582	0.584	0.489	0.419
Lubuskie	0.152	0.142	0.114	0.120	0.092	0.105	0.109	0.106
Wielkopolskie	0.388	0.345	0.275	0.276	0.266	0.265	0.222	0.208
Zachodniopomorskie	0.432	0.376	0.321	0.290	0.254	0.260	0.236	0.226
Dolnośląskie	0.354	0.306	0.287	0.272	0.238	0.248	0.263	0.266
Opolskie	0.961	0.964	0.922	0.864	0.750	0.796	0.705	0.672
Kujawsko-pomorskie	0.324	0.290	0.249	0.243	0.222	0.244	0.222	0.199
Pomorskie	0.166	0.158	0.135	0.136	0.127	0.140	0.136	0.130
Warmińsko-mazurskie	0.108	0.096	0.079	0.079	0.072	0.065	0.058	0.060

Sources: Bwanakare S., based on Eurostat and Polish Statistical Office (GUS) data.

the row totals and the column totals. The information on the average weight of each regional greenhouse emission comes from the last column, each element of which is divided by the total of last column totals. Then, using regional weights from the column totals, we have set up the prior matrix of cross-regional greenhouse emission per industrial sector, the total of which must sum up to the row totals.

Such a prior matrix reflects uniform probability distribution of all regional greenhouse emission within all industrial sectors. As noted in the above section, besides classical normality and consistency conditions introduced in the model (Eqs. (6)–(10)), the Kuznets economic variable $C_{k\bullet}$ has been introduced in the model as new statistical evidence in the sense of the Bayesian optimal rule. Table III presents the empirical values of

the above Kuznets variable over the period 1998–2005. These computed values are not deterministic since neither the quantity of greenhouse pollutants nor the GDP are known with certainty. Thus, we treat them as random variables from which definitive posterior values will result through model optimization. In theory, other things remaining equal, one expects the contemporary value in-

formation (i.e., from 2005) on the $C_{k\bullet}$ should best discriminate the prior, i.e., the initial information of Table III, in favour of the unknown, to be estimated, true parameter values of the regional greenhouse pollutants emission. In reverse, information conveyed by the $C_{k\bullet}$ for the year 1998 should least impact the contemporary model.

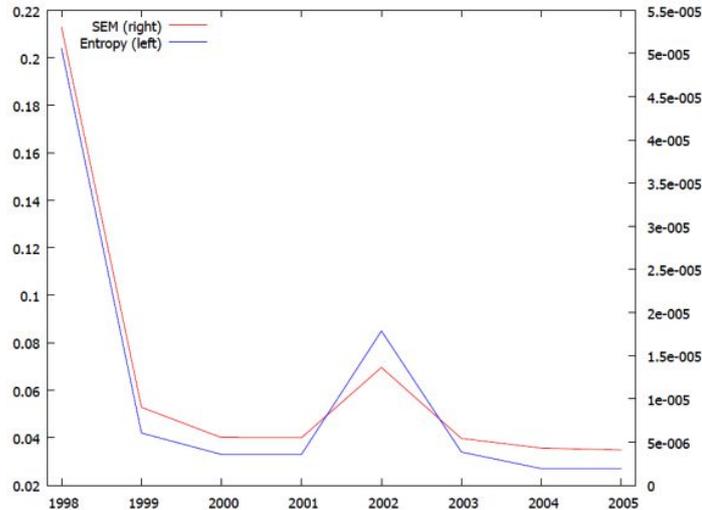


Fig. 1. Average error discrepancy between the prior and the post entropy posterior coefficients as a function of the prior periodical coefficients of greenhouse pollutants tons per one unit of GDP.

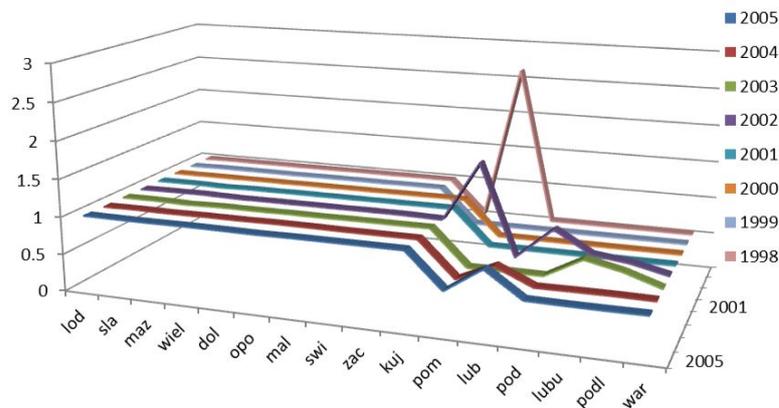


Fig. 2. Evolution of posterior regional coefficients of greenhouse pollutants tons per one million unit of GDP over the period 1998–2005.

As a result of a simulation using different $C_{k\bullet}$ values for each period, 1998–2005, Fig. 1 plots standard errors between the prior cross-regional industrial sector greenhouse emission and the post-entropy outputs. As expected, we observe that, in general, the average error steadily increases in proportion to the distance in time from the most current year 2005. This is confirmed by the same Fig. 1, which displays the different values (of the period 1998–2005) of the entropy coefficient change expressed in Eq. (7). In this case, the entropy value is

smaller when we take the coefficients from the period 2005. In the same figure, the coefficients for the period 2002 seem to produce higher standard error or entropy.

The explanation remains unclear even if a plausible reason could dwell in the quality of statistical data. Figure 2 shows the dynamic evolution of posterior regional coefficients of greenhouse pollutant tons per one million units of GDP over the period 1998–2005. They seem more stable and closer to each other over time than the priors

(see Table III). Outputs suggest two categories of Polish province: 10 with higher greenhouse pollutants per unit of GDP and 6[§] with lower coefficients. The interregional coefficients are, over time, quite similar in both regional groups. This could be related to comparable technology and economic welfare within the regions involved. In the same Fig. 2, there are some outstanding values of $C_{k\bullet}$ in regions with lower coefficients. Except the observation value of the year 1998 related to the province Pomorskie, most of these values concern 2002. For that year, GAMS-generated outputs have not reached the optimum. It is plausible that original data may contain some inconsistencies with respect to the DGS.

The non-reported, here, post-non-extensive entropy outputs seem to confirm the known tendency of Gibbs–Shannon entropy in which smaller probabilities are shrunk and higher probabilities, in reverse, dominate in the space of solutions[¶], suggesting that Tsallis entropy should belong to the Stein estimator family [8].

Thus, as expected, the above simulations tend to prove that additional data, consistent with the model, will reduce the entropy and improve the obtained results.

4. Concluding remarks

As we know, adding a new piece of a priori information consistent with the properties of the presented model will significantly improve the quality of prediction of the emitted greenhouse pollutant quantities. This has been demonstrated in this paper through information related to Kuznet’s coefficients $C_{k\bullet}$ in Table II, where recent information tends to reduce the model error better while older information does it less accurately. Interestingly enough, as displayed in Fig. 1, except for the year 2002, information conveyed by $C_{k\bullet}$ in Table II over the period 1999–2005 remains very stable with respect to constraint impact on criterion function and impacts the level of model prediction. Such a result where observation corroborates theory — plus the relative stability of the optimal outputs with respect to new environmental evidence embodied within the time-related values of Kuznet’s coefficients — should be treated as a good sign that the optimal model outputs obtained might be somewhat closer to the true, unknown pollutant emissions across industries and regions. Another relevant point is that the coefficients $C_{k\bullet}$ stand for the main random determinant of differences between data from GUS (total last column) with those from Eurostat (total last row). It is worth adding that all the optimal solutions obtained correspond to the Tsallis q-parameter — treated in the model as a variable, equal to one. Another finding to be noted, as has been observed in the case of forecasting input output systems through Gibbs–Shannon en-

tropy [8], Tsallis entropy estimators belong too to the family of Stein estimators. As such, smaller probabilities are shrunk and higher probabilities then dominate in the solution space. Fortunately, adding more a priori information in the model will enhance parameter precision and then allow for recovering the influence of smaller events while dampening that of bigger events.

References

- [1] A. Dragulescu, V.M. Yakovenko, *Physica A* **299**, 213 (2001).
- [2] R. Rak, J. Kwapien, S. Drozd, *Physica A* **374**, 315 (2007).
- [3] F. Nielsen, R. Nock, *J. Physics A-Math. Theor.* **45**, 032003 (2012).
- [4] A.N. Tikhonov, V.I. Arsenin, *Solutions of Ill-conditioned Problems*, John Wiley & Sons, 1977.
- [5] E.T. Jaynes, *Probability Theory: The Logic Of Science*, Washington University, USA 1994.
- [6] C.E. Shannon, *AT&T Tech. J.* **27**, 379, 623 (1948).
- [7] S. Bwanakare, *Entropy* **16**, 2713 (2014).
- [8] A. Golan, G. Judge, D. Miller, *Maximum Entropy Econometrics: Robust Estimation with Limited Data*, Wiley in Chichester, England, 1996.
- [9] S. Robinson, A. Cattaneo, M. El-Said, *Econ. Systems Res.* **13**, 47 (2001).
- [10] F.H. de Campos Velho, E.H. Shiguemori, F.M. Ramos, J.C. Carvalho, *Comput. Appl. Math.* **25**, 307 (2006).
- [11] X. Gabaix, *Power laws in economics and finance*, September 2008, <http://www.nber.org/papers/w14299.pdf>.
- [12] C. Tsallis, R.S. Mendes, A.R. Plastino, *The Role of Constraints within Generalized Non-extensive Statistics in: Physica A: Statistical Mechanics and its Applications*, North-Holland, 1998.
- [13] R. C. Venkatesan, A. Plastino, [arXiv:1102.1025v3](https://arxiv.org/abs/1102.1025v3), 2011.
- [14] S. Kullback, *Information theory and statistics*, John Wiley and Sons, NY 1959.
- [15] L. Illge, R. Schwarze, *A Matter of Opinion: How Ecological and Neoclassical Environmental Economists Think about Sustainability and Economics*, German Institute for Economic Research, 2006.
- [16] D. Runnals, *S.A.P.I.E.N.S.* **4**, (2011).
- [17] N. Hanley, J. Shogren, B. White, *Environmental Economics in Theory and Practice*, Palgrave, London 2007.
- [18] E.Z. Shen, J.M. Perloff, *J. Econometrics* **104**, 289 (2001).
- [19] F. Pukelsheim, *JASA* **48**, 88 (1994).
- [20] S. Abe, G.B. Bagci, [arXiv:cond-mat/0404253](https://arxiv.org/abs/cond-mat/0404253), 2004.
- [21] C. Tsallis, *Introduction to Non-extensive Statistical Mechanics: Approaching a Complex World*, Springer, Berlin 2009.
- [22] S. Bwanakare, *Acta Phys. Pol. A*, **123**, 502 (2013).
- [23] S. Bwanakare, *Acta Phys. Pol. A* **117**, 647 (2010).
- [24] Eurostat, <http://epp.eurostat.ec.europa.eu/portal/page/portal/statistics/themes>.
- [25] GUS, http://stat.gov.pl/bdl/app/dane_podgrup.dims?p_id=528516&p_token=0.9178699365269101.

[§]These are: Pomorskie, Lubelskie, Podkarpackie, Lubuskie, Podlaskie, and Warminsko-Mazurskie.

[¶]For the demonstration of this attribute, see, e.g., Golan et al [8]