Working paper No. 3 (c) (2000)

PRELIMINARY OPTIONS FOR METHODOLOGIES TO APPLY ADJUSTMENTS UNDER ARTICLE 5.2 OF THE KYOTO PROTOCOL

Fuel combustion activities

Interpolation, extrapolation and approximation using drivers

Expert report

prepared for the

UNFCCC Secretariat

20 February 2000

Contents

This report was prepared by
Miloš Tichý, Tristolicna 5, 150 00 Prague 5, Czech Republic
Phone/fax : #420 2 2162 4740, email: milos.tichy@sujb.cz

1 Introduction

A technical estimate (TE) of a GHG emission, for the purpose of this paper, is understood as a calculation which may replace a reported emission value in the inventory which is incomplete or not performed in good practice way. This may concern emissions both in the past and in future. The TE can be initiated by two different findings:

- A country itself learned that there are no available data for an inventory conform with the IPCC guidelines and good practice for a certain source(s) even on the ground of the simplest way (Tier1). Therefore a TE could be used to complete the inventory. This may not be the case of the inventory for the last period but more probably it is the case of older data.
- An inventory review team has found that the inventory is incomplete or performed in a way, which differs from IPCC guidelines or good practice.

These two cases differ in existence or availability of appropriate[1] data, which methodologically important for the crucial question: *Can a performed TE be verified?* In the first case absence of the appropriate data is evident and the TE cannot be verified, whereas in the second case missing data can be found and compared with the TE.

Whereas in a reality both situations (verifiable and non-verifiable results) may appear, a method for the TE will be demonstrated here on verifiable data i.e. estimated data item will be "forgotten" for a moment and "re-found" data item will be used for verification of the TE result at the end.

A TE can be also used for examination of submitted data. This is understood as a procedure roughly described above: Apparently complete (no number is missing) inventory is presented by a country to the in-depth review team. There is no sign of using procedures out of good practice but the team may conclude that some data have a lower reliability e.g. time series have no smooth course. The team may perform a TE to get a comparable number to the worrying data item. This is a procedure with a lower validity than verification or checking but it is usually faster and it does not require additional information from presenting country.

The following paper deals with one group of possible methods for TE more than executing a TE for concrete data. The methods of interpolation and extrapolation of data for fuel combustion sector are explained and their possibilities are demonstrated on examples. For these methods, correct emissions have to be known for at least two years.

Fuel combustion sector is the most important sector for most of Annex B countries. It covers 75-95 % of their total emission. But it is also the simplest sector from methodological point of view in the following sense: broadly spoken, that all carbon burned is released and examination of fuel carbon content is well known task. The exception is 1-3%[2] of carbon stored in black and ash.

The emissions are calculated as a product of activities and emission factors. Activities mean apparent spent fuel consumptions and they are relatively well-documented values. Emission factors are closely related to carbon content as follows from above and with the calorific value if the amount of spent fuel is reported in its heat content.

The main assumption used for the method(s) presented here is that both activities and emission factors are quantities with smooth course in time. This follows from smooth change of fuel consumption and structure in the case of activities. In case of emission factors smooth changes result from small changes of fuel chemical composition in one mine (or oil well). This assumption is of course more satisfied on more aggregated levels of the inventory.

[1] By the term appropriate data we understand data for an inventory to be complete and performed in conformance with good practice .

[2] This is without extensive measurement almost conventional value.

2 Overview of interpolation and extrapolation methods

Interpolation and extrapolation is a task to find a function value $f(x)$ which is unknown because the original function is known only as a table for discrete argument values: pairs $(x_1,y_1),\ldots, (x_n,y_n)$, where is, $y_i=f(x_i)$, $i= 1,\ldots,n$.

The task is solved by an approximation of a function $f(x)$ by another function $P(x)$. Generally two techniques are available for approximation: interpolation and regression. Their main difference is in principal condition for approximation function:

- Equivalence of function values $f(x_i)= P(x_i)$ for all known values $f(x_i)$ is required in the case of interpolation.
- Regression is based on a softer condition: the difference of original and approximate function is minimized. The most frequent functional minimized is sum of squares $(f(x_i) - P(x_i))^{2.}$.

Interpolation is used for data, which are precisely known (no uncertainty) and they are not very much scattered. **Regression** is frequently used for randomly scattered experimental data when this (random) scatter should be filtered out or when a parameter of a predefined formula should be assessed. Both methods (techniques) use the same formula, for both possible cases: an argument of unknown function value x is inside (interpolation) or outside (extrapolation) of the set $x_1,\ldots,x_n$,

Terminological note.

"Interpolation" is used in this context in two meanings:

- **a task** to find an approximation of a missing function value for an argument which *inside* of a set of known values (opposite is extrapolation)
- **a method or technique** to solve the interpolation or extrapolation task with condition $f(x_i)= P(x_i)$.

Without introduction of a new terminology it is impossible to separate the two meanings and the right sense should be understand from the context.

The approximation function $P(x)$ is in most cases a polynomial $P_m(x)$.

Interpolation *technique* is an approximation by a polynomial function which coefficients are calculated to fulfil the condition $f(x_i)= P(x_i)$. One of possible formulas used for calculation of polynomial coefficients is in Annex 1. Accuracy[3] of the technique depends on the order of polynomial which defined by number of known function points around x. The most usual is the linear interpolation when a function value $f(x)$ is calculated from its neighbouring values. Higher order polynomials are calculated using more points, which can be on "one side" of the approximated value. An example is provided in Annex 1.

Regression is a substantially more complicated procedure. The condition of equivalence of function values is replaced by minimization squares of differences. More detailed description is in Annex 2.

It is necessary to have in mind that the above methods (interpolation and regression) provide almost uncountable amount of possibilities. For both methods the number and the selection of known function values (points) is extremely important. In case of interpolation the number of points taken into account is given by selected order of polynomial (or vice versa). For regression method these two parameters (number of known points and polynomial order) are selected almost independently.

Commercially available software offers substantial packages for interpolation and regression. Microsoft Excel[4] offers the following tools:

- For use of interpolation techniques spectrum of ready functions is limited:

[3] Accuracy in this paper is the difference between the original function and an approximation. This may be inconsistent with statistical definitions.

[4] The reason for selecting Excel is vide availability and this software was used for "IPCC software for GHG Inventories".

- For liner interpolation functions identifying position of an argument x in the array $x_1,\ldots,x_n$.[5] can be used and interpolation formulas similar to those in Annex 1 can be developed.
- Cubic interpolation can be performed using an algorithm from [2].

- Spectrum of tools usable for regression is wider:
  - A simple Excel function LINREGRES provides estimation of parameters of a polynomial approximation of any reasonable order (not only linear as the name may induce).
  - Graph facility and Analytical tools offer additional possibilities: regression by most frequent functions (including polynomials, exponential) or regression by a user defined function.

These tools are used in the following examples.

3 Practical application

The following part contains a set of examples how the methods and techniques described above can be used in replacing of missing values in emission time series. The first part concerns application of interpolation and regression techniques and in the second one the utilization of a driver or index. It employs the same mathematical technique of regression.

As mentioned above both possibilities for TEs have a large number of variations and a complete and comprehensive description is impossible at least in the frame of this study. Therefore a set of examples was selected to illustrate possibilities and limitations the above methods and techniques.

Procedures which are described in this part are based on intuitive knowledge that all quantities entering the inventory either directly (emission, activity, emission factor) either indirectly as drivers[6] have smooth course (relatively small changes in time) unless a substantial condition is changed.

3.1 Interpolation and extrapolation techniques

3.1.1 Interpolation of emission value

Time series of relative total emission from fuel combustion [3] were taken as the first example. Emission for the year 1994 was deleted to model a missing data item and an approximation was calculated using interpolation and regression techniques. Seven approximations that differ in number of points taken into account and in order of a polynomial were tested. Results are in Tab. 1 and Fig. 1.

The emission for 1994 year is the lowest value of the usual U shape of emissions of Central and East European countries. Therefore for all variants, interpolation overestimates the true function value because a small valley is not fully taken into account. As expected increasing order of polynomial i.e. taking into account of more information on emission before and after the missing year brings a better result in some cases. Second order polynomial (quadratic) takes into account three points: two before and one after the approximated point and vice versa. The difference is substantial (see Tab. 1): if the third point is chosen from older values (1992 year) a much worse result is obtained compared to the choice of emissions for years 1993, 95 and 96. The reason is apparent from the graph (Fig.1). This example demonstrates sensitivity of the TE on “subjective” selection of input information. An increase of polynomial order does not bring a better result – this is the limit of the technique.

Regression in this example takes into account all available data (but choice of any subset is possible). Therefore result from linear regression is worse than linear interpolation – approximation of a U shape by line is not very advantageous. For higher polynomial orders much better results were obtained.

An additional test was performed with data from [4]. Whereas the previous data set contained 6 points the IEA data [4] has about 25 points. No interpolation technique was used because it would not have provided new information. In Tab. 2 an extremely bad and expected result comes from linear regression but comparable results are for higher orders. It is apparent that there is no reason to assess higher orders than 3rd.

[5] Interpolation formula in [2] differs from formula in Annex 1 but they are equivalent.

[6] For explanation of a driver see later.

It is apparent that best accuracy is a fraction of percent and the worst is about 13%. In a real situation nobody knows that the missing value is the lowest value of the U shape or emission is just in line with the previous ones and then all value have the same probability i.e. linear interpolation as the easiest technique is preferable. Regression may generate better results if a shape of the curve is clear for longer time series and therefore some values out of the straight line are more probable.

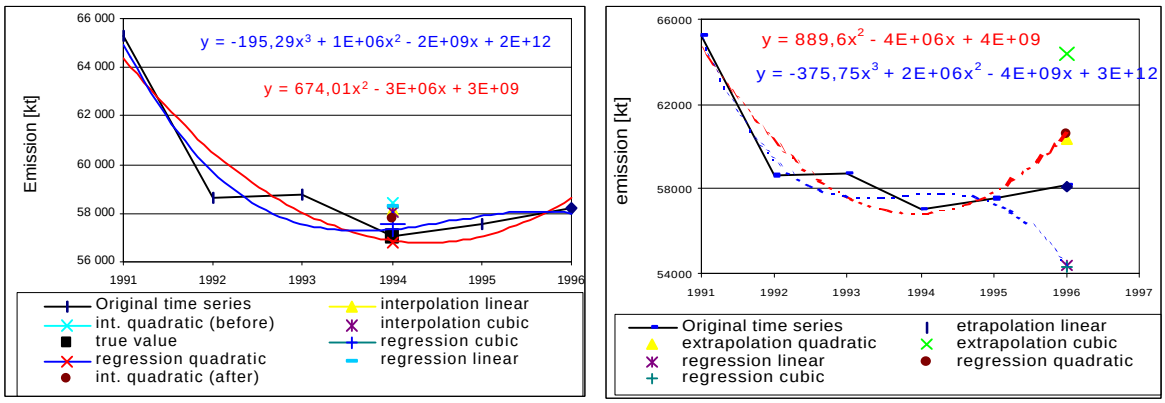
Tab. 1 Accuracy of different approximations of missing one year emission

Approximation	Accuracy		
	Interpolation-short series (data from Hungary)	Interpolation-longer series (IEA data)	Extrapolation (data from Hungary)
Interpolation linear	1,95%		-0,15%
quadratic (before)	2,37%		3,68%
quadratic (after)	1,25%		
cubic	1,81%		10,66%
Regression linear	2,17%	13,01%	-6,55%
quadratic	-0,50%	-1,56%	4,15%
cubic	0,94%	-0,52%	-6,70%
4th order		-0,61%	

3.1.2 Extrapolation using previous values

The extrapolation example is very similar to interpolation – approximation function is selected and missing value calculated. The same data set was selected to demonstrate available techniques. Emission of 1996 year was "forgotten" and different approximations calculated. Results are inTab.2 and Fig. 1.

Fig. 1: Approximation of a missing emission value



The emission time series is almost linear in two last periods (1994-96). Therefore the simplest linear extrapolation resulted in the best TE. In contrast to previous example (interpolation) higher orders of approximation (both interpolation and regression) are worse with increasing order. If the two techniques employing polynomial of the same order are compared, linear regression is very inaccurate because it takes into account all known points (the same reason as for interpolation). A huge inaccuracy harms also extrapolations of higher orders for both techniques, which takes into account more known points. Intuitively we can deduce that a long time series approximated by regression of the second or third order may bring better results than the same technique which takes into account five points only. All cases form a certain range of reachable accuracy: from almost zero to 10%.

Impossibility to derive guidelines for technique selection and a choice of number of points to be taken into consideration for all cases can be concluded. But general guidance can be derived as follows:

- Use of polynomials of higher orders (3-4) may bring better results for longer time series and regression technique.
- Inclusion of more points for interpolation technique may not bring substantially better result.

- More points taken into account bring more reliable results for slowly changing functions.
- Regression technique seems to be more reliable and more objective than interpolation. It may bring a more reliable (than interpolation) result at least for “randomly” scattered data.

3.2 Approximation using drivers.

In contrast to previous methods, which were based purely on observation of shape of quantity time series, the approximation using drivers is based on a deeper knowledge of ”driving forces” in the energy economy system. Missing points in time series of emission are filled by values calculated as a simple function (most likely as a multiple) of an indicator of driving forces, called driver, which is assumed available in a complete time series. Any quantity, broadly spoken, can be used as a driver but the function (relation driver -> emission) should be:

- as simple as possible,
- time independent i.e. possible parameters of the function should not change in time and
- causal[7].

A causal relation can be found at least for indicators of economic activity and an activity in an inventory sense. Assuming approximately constant emission factors, also the emission can be driven by the same driver. The driver is usually a well statistically documented quantity published in long and complete time series. Probably the most frequently used driver is the GDP or its parts relevant to specific sector.

Examples of drivers and driven functions can be arranged into the following causal chains:

- economic activity (GDP) -> production of electricity -> activity in 1A1 sector -> emission in 1A1 sector,
- GDP in production sector -> emission in Industry (1A2)
- Population -> emission in Residential sector (1A4b).

A limited causality[8] of the chains above is apparent. To prove the relations, the data from IEA [4] and UNFCCC [6] were used. The first source is the only available data set, elaborated by the same methodology, which contains all kinds of data ranging from economic data, energy balance to CO_2 emission in time series since 1960. The other data were provided by countries to the UNFCCC and they cover only emissions for period 1990-97. There is a difference from IEA data but it was considered negligible for the purpose of demonstration of approximation methods. Therefore the first database will be used further.

The approximation method was tested on a group of countries (USA, UK, France, Germany, Poland, Hungary). For some driver -> quantity relations only a subgroup was used when the full-scale test was considered not necessary. Therefore the presented analysis may serve as a comprehensive set of examples but it is not in any case a complete analysis. Such analysis has to cover not only many more countries but also more sectoral and sub-sectoral emissions.

[7] The causal relation should be based on knowledge of internal mechanism and be proved (at least illustrated) to avoid random coincidence. Technological energy-economy models are based on the described method.

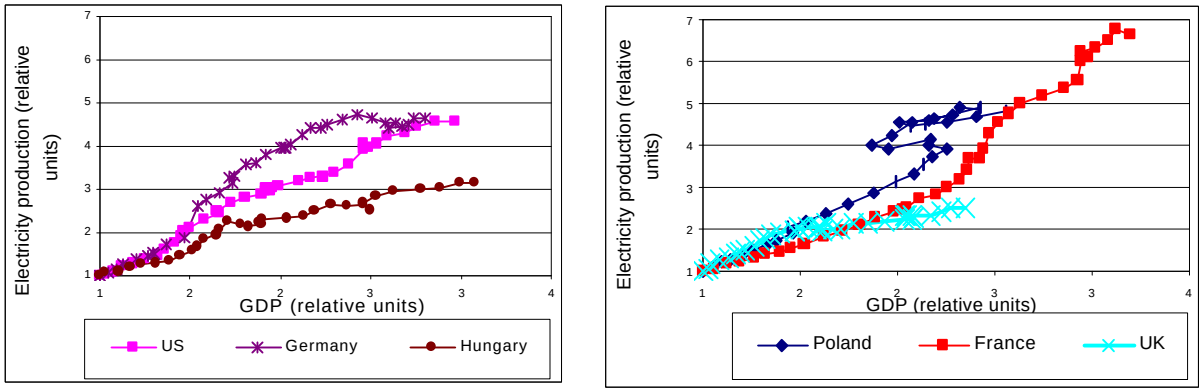
[8] Limitation of causality means that other influences have to be taken into account also:

- Modern economies have lower growth of electricity consumption than GDP growth (but this difference may be constant in time).
- Structure of fuels (and then emission factors) for electricity production is changing. This influence is substantial in small economies especially when a new non-fossil source is introduced.
- Differences in sectoral boundaries: electricity for industry enterprise own production is attributed in IPCC inventory methodology to sector of industry whereas in other methodologies (e.g. IEA) it belongs to transformation sector.

3.2.1 GDP -> Electricity production/consumption

For the purpose of this study as an illustration of driver, relation of GDP and electricity production[9] was selected (see Fig. 2). For US, France and Hungary the relation is almost proportional (with less accuracy for the second two). For Poland, Germany and the UK two coefficients for two different periods can be found[10]. This relation is only the first step of the chain mentioned above ending by emissions and can be used only when the TE covers calculation of activity only having emission factor from another source.

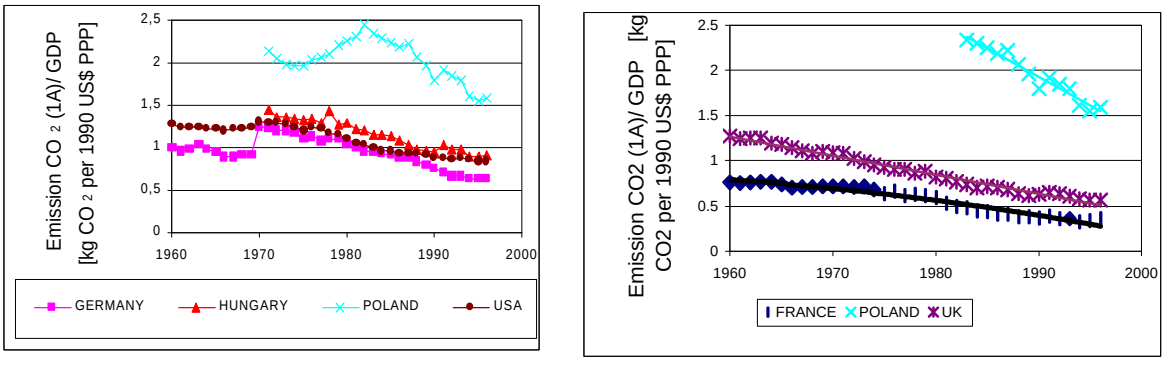
Fig. 2: Relation of the total electricity production and the GDP. Both quantities are plotted as relative to the value for 1960 year (1971 for Hungary and Poland).



3.2.2 GDP-> Total Emission from Fuel Consumption

Frequently used indicator "emission per unit of GDP" can be used to assess a driving relation GDP -> emission. Total fuel combustion CO_2 emissions (sector 1A) per USD of GDP (of 1990 year) are plotted in Figs. 3 and 4. No time-constant coefficient relating GDP and emission can be found i.e. there is no a simple proportional relation between this two quantities. If the second requirement (the time independence) is given up, a possible solution appeared: fitting (regression) the ratio by a polynomial. Both chance and limitation to keep the relation simple are apparent from in Fig. 4. Data for the USA may serve as example of well fitted linear relation, for Poland the same can be used for a limited period (1982-96) and with limited accuracy. More oscillating curve for France can be fitted by a polynomial of the second order.

Fig. 3: Emission from fuel combustion per unit of GDP



[9] In this context production or consumption of electricity can be used interchangeably because net import for the countries examined (UK, US, Germany, France, Hungary and Poland) is less than 1% of the production except few years when Hungary imported about 3,5% of electricity production.

[10] Reasons for this can be found also: Germany has two different statistical periods – West Germany until 1969 and the sum of West and East Germany since then (i.e. ten years before re-unification), for Poland is the

Therefore the emission in sector, 1A can be calculated using the following equations:

$$E_{1A-Poland} = (-0{,}0636 * t + 128{,}55) * GDP$$

$$E_{1A-UK} = (-0{,}021 * t + 42{,}464) * GDP$$

$$E_{1A-France} = (-0{,}0002 * t^2 + 0{,}7193 * t - 696{,}81) * GDP$$

where *t* is year and GDP is GDP expressed in USD (constant 1990) of PPP. The equations can be used not only for TE for points inside of the time range i.e. as interpolation using regression but also for points outside the range near of the limits i.e. for extrapolation.

The accuracy of a TE of a single emission value missing in the series[11] is few percent for most of points inside the time period (1960-97). For TE at more "outlier" points accuracy is not worse than 15%.

3.2.3 EMBEDFossil electricity production -> Emission

To get driving relation for sectoral emission e.g. emissions from industry (1A2) or residential (1A4b) more specific drivers should be identified. Much closer causal relation can be build between production of electricity in public power/CHP stations and corresponding emissions. Their ratio is the widely used indicator "CO_2 efficiency" of fossil electricity production. To avoid the influence of "non-emitting" technologies (hydro, nuclear), electricity production from fossil sources is used in denominator of the ratio only. Time series of the indicator is plotted in Fig. 5 for France, Germany and Hungary.

In spite of very close relation (activity -> emission) the possibilities are limited. A simple proportional time-independent ratio can be found for a limited period 1973-90, which is not the most interesting time period and for Germany only. Changes of the indicator in the period of UNFCCC interest (since 1990) reflect changes in fuel mix.[12] Time dependent relations similar to those for GDP -> emission from 1A1a can be obtained as follows:

$$E_{1A1a-France} = (3.*10^{7} t^4 - 0.0027 t^3 + 8.3589 t^2 - 11300 t + 1.*10^{6}) * P$$

$$E_{1A1a-Germany} = (1.*10^{-6} t^4 - 0.0092 t^3 + 27.386 t^2 - 36326 t + 2.*10^{7}) * P$$

$$E_{1A1a-Hungary} = (-6.*10^{-6} t^4 + 0.0476 t^3 - 140.66 t^2 + 184753 t - 9.*10^{7}) * P$$

where *P* is the public electricity production and *t* is the year[13]. Extensive testing of relation form and order versus accuracy was preformed. Accuracy reached for polynomial of the fourth order was generally the same as in chapter 3.2.2: better than 5% form most of points and 10-15% for remaining ones. For polynomial of the third order result almost the same but the second order they were substantially worse.

---

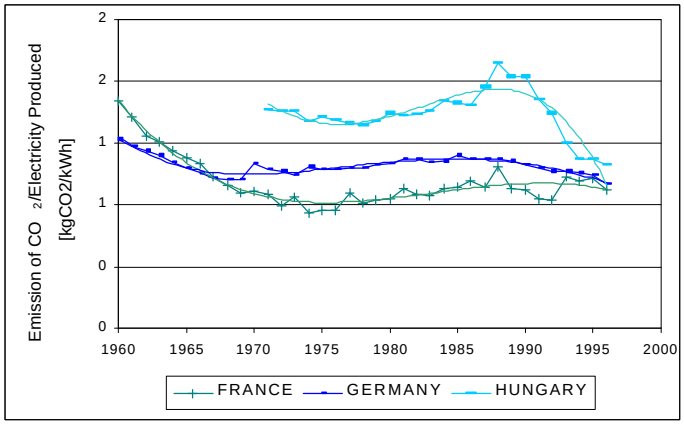
reason restructuring economy and the UK undergone a gradual but substantial change of fuels used for electricity production since late 70th .

[11] The same example was performed as for interpolation: one value is "forgotten" and then used for verification,

[12] Definition of public producer and autoproducer which is also not very clear and is changing in transforming economies may influence scatter of curves in Fig. 5.

[13] Note that coefficients are substantially rounded.

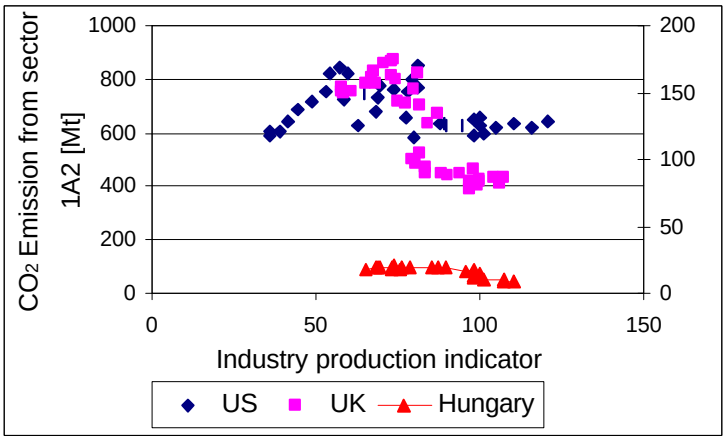
Fig. 5: "CO_2 efficiency" of public electricity production, CO_2 emitted per kWh electricity produced from fossil sources.



3.2.4 Industry production -> Emission from Industry

More serious problem arises in establishment of a driver for emission in industry. IEA industry indicator has been chosen and its relation to emission in industry sector (1A2) is in Fig. 6. In this case the approximation may yield into TE with accuracy lower than 20%.

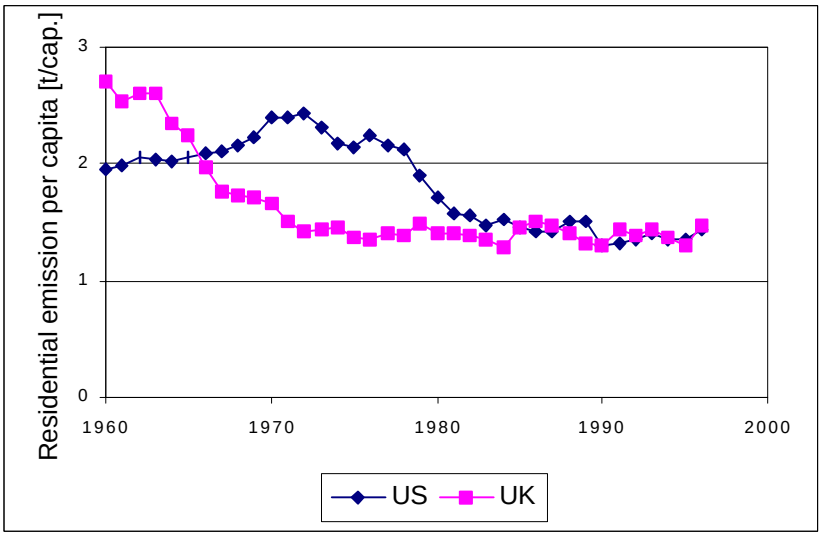
Fig. 6: Relation of industrial production to corresponding emission



3.2.5 Population -> Emission from Residential

Population can be an assumed driver for emission from residential sector. The data for USA and UK were tested (see Fig. 7). Emissions are proportional at lest for limited time period. For the period since 1973 (for the UK) and since 1983 (for the US) the ratio is almost constant and equal for both countries (1.4 t_{CO2}/cap.). Accuracy of approximation is almost the same as for relation GDP-> Emission 1A i.e. few percent for most of points and below 15% for the rest of emissions.

Fig. 7: Relation of emission and population: emission from residential sector per capita



4 Conclusions

1. A missing emission value in a time series can be replaced by a value of an approximation function (the task is called Technical Estimation -TE) gained by one of inter/extrapolation techniques mentioned in the next paragraph or calculated from the value of a corresponding driver for the same year when an appropriate driver is found.
2. A short theoretical background of a TE (i.e. replacement of a missing value of an emission) based on approximation by a simple polynomial function was given. Two techniques can be employed to obtain the approximation function: inter/extrapolation in a usual sense and regression. No uncertainty of emissions is assumed here, but the methods allow an extension, which can yield into estimated uncertainty of the TE.
3. An approximation by a polynomial (gained by either interpolation or regression) is relatively straightforward. The accuracy depends on scatter of the data: probability of bigger errors increases for more scattered data. This can be managed up to some extend by selection of polynomial order and number of points influencing the approximation.
4. The difference between a TE and the right value is called accuracy[14] in this paper. The accuracy of a TE was assessed in testing examples but in a "real" situation is unknown (for an idea of accuracy indicator see later paragraph). For tested cases it was in order of few percent (<5%) for interpolation using either an interpolation formula or regression. For extrapolation (i.e. a TE for point outside of known time-series) using the same methods the accuracy is slightly lower. The regression technique can be slightly preferred especially when longer time series is available.
5. A TE using a driver requires a substantial effort to find a proper driver and to establish reliable relation driver -> emission. For some cases, where a relation driver emissions was expected, this relation was found to be too complex and unreliable, so that the method failed. Drivers for very aggregated emission (1^{st} and 2^{nd} levels of IPCC96) can be found easier because general indicators of economic activity can be used. Similar indicators can correspond also to emissions on lower disaggregation levels but an effort to obtain their time series can be comparable to get the right data (emissions) from national sources.
6. A substantial effort was devoted to identify drivers and to specify their relation to emissions from fuel combustion where a promising driver was identified. In four cases (GDP ->Emission in sector 1A, Public Electricity production -> Emission in sector 1A1a, Population -> Emission in sector 1A4b and GDP -> electricity production) driven quantity was found to be proportional to the driver in a limited time range or their relation can be approximated by a simple polynomial of a low order. The relation of a possible driver "Industry activity" to corresponding emission quantities appeared to be too complicated to be used for the purpose of a TE. The presented selection may be taken as incomplete examples. Identification and specification of such relations will be more complicated for more specific sectors as mentioned in point 5.
7. Experience is needed to perform a good TE using both presented methods. Different subjective expectation of emission trends expressed by different experts (especially in extrapolation) may yield different results. The presented methods cannot be developed into guidelines valid for all countries, emissions and time periods. Therefore TEs should be solved case to case by experienced personnel. To perform a TE which are independent on person and for which experience is not required is possible in "trivial" cases i.e. inter/extrapolation of a quantity having linear course or when accuracy of a TE between two known points not better than around 5% is acceptable.
8. General recommendation for method selection: a TE based on polynomial approximation is possible either by classical inter/extrapolation formulas or by regression especially for more aggregated and therefore less scattered data. These methods can be used also for less aggregated data but the accuracy may decrease. To perform a TE using a driver is more time consuming, expected uncertainty may be larger and for some quantities and periods may fail.
9. An assessment of TE quality in case of missing value is impossible. But when regression is used an average of existing data from the approximation can be used as an indicator of the accuracy.

[14] Accuracy is not a statistical term in this paper.

10. All the tasks using the above mentioned methods can be solved a by commercial spreadsheet software. Development of macrocodes can be advisable for methods using interpolation formula(s) or in case of numerous TEs.

5 References

[1] Rektorys: Handbook of Applied Mathematics (in Czech)

[2] Urbánek T, Škárka J.: Excel 97 for scientists and technicians, Computer Press, Prague 1998, in Czech

[3] Emission data for Hungary, data received from the FCCC

[4] IEA database containing economy indicators, energy balances and CO2 emissions.

[6] Emission data base of the UNFCCC accessible via the internet

Note: references [1] and [2] are in Czech but the first is a general mathematical reference book and the second is a general guide for using MSExcel. Both books can be easily replaced by some English ones.

6 Annex 1

General interpolation (Newton) formula is as follows:

$$P_m(x) = f(x_o) + (x - x_o) * (f(x_o, x_1) + (x - x_1) * (f(x_o, x_1, x_2) + \ldots + (x - x_{m-1}) * (f(x_o, x_1, \ldots, x_m)))$$

where $f(x_o, x_1, \ldots, x_j)$ is the relative differences of j-th order. Definition of relative difference is apparent from Tab. 1:

Tab. 1

argument	function value	Relative difference		
		1st order	2nd order	3rd order
x_o	$f(x_o)$	$f(x_o, x_1) = \frac{f(x_1) - f(x_o)}{x_1 - x_2}$	$f(x_o, x_1, x_2) = = \frac{f(x_1, x_2) - f(x_o, x_1)}{x_2 - x_o}$	$f(x_1, x_2, x_3, x_4) = = \frac{f(x_1, x_2, x_3) - f(x_0, x_1, x_2)}{x_3 - x_1}$
x_1	$f(x_1)$	$f(x_1, x_2) = \frac{f(x_2) - f(x_1)}{x_2 - x_1}$	$f(x_1, x_2, x_3) = = \frac{f(x_2, x_3) - f(x_1, x_2)}{x_3 - x_1}$	
x_2	$f(x_2)$	$f(x_2, x_3) = \frac{f(x_3) - f(x_2)}{x_3 - x_2}$		
x_3	$f(x_3)$			

Accuracy increasing with polynomial order, for a simple function, can be demonstrated by the approximation of sin(x) by polynomials through four known points of sin(x) (0^o, 30^o, 45^o, 60^o, 90^o). The polynomials can be used to approximate the value of sin (35^o) by interpolation and sin (95^o) by extrapolation[15] as presented in Tab. 1.

Tab. 1 Interpolation of value of sin 35^o and extrapolation of value of sin 95^ousing using different orders of approximation polynomial.

approximation order	**sin 35^o**		**sin 95^o**	
	value	accuracy	value	accuracy
true value	0,5736	0,00%	0,9962	0,00%
1.order	0,5690	-0,79%	1,3975	40,28%
2.order	0,5722	-0,24%	1,1909	19,55%
3.order	0,5735	-0,016%	0,9669	-2,941%
4.order	0,5736	0,004%	0,9958	-0,043%

Tab. 2: Details of inter/extrapolation example.

x[deg]	f(x)=sin x	$f(x_1,x_2)$	$f(x_1,x_2,x_3)$	$f(x_1,x_2,x_3,x_4)$	$f(x_1,x_2,x_3,x_4, x_5)$	**35^o**	approx.	accuracy
0	0,000	0,0167	-6,355E-05	-7,257E-07	2,672E-09	true value	0,5736	0,00%

[15] Extrapolation is, rigorously taken, an "incorrect" operation i.e. function value is estimated in a an argument range where no-information is available.

30	0,500	0,0138	-1,071E-04	-4,852E-07		1.order	0,5690	-0,79%
45	0,707	0,0106	-1,362E-04			2.order	0,5722	-0,24%
60	0,866	0,0045				3.order	0,5735	-0,016%
90	1,000					4.order	0,5736	0,004%

7 Annex 2

Regression is generally a task to find parameters $z_1,\ldots,z_k$ of a function $P^{z_1,\ldots,z_k}(x)$ so that

$$\sum_{i=1}^{n}\left[f(x_i)-P^{z_1,\ldots,z_k}(x_i)\right]\rightarrow \min.$$

The final formulas for parameters $z_1,\ldots,z_k$ calculated as functions of $x_1,\ldots,x_n$ depend on formula of $P^{z_1,\ldots,z_k}(x)$. For first order polynomial $P^{a,b}(x)=ax+b$ we will get

$$b=\frac{\sum_{i=1}^{n}x_iy_i-\frac{1}{n}\sum_{i=1}^{n}x_i\sum_{i=1}^{n}y_i}{\sum_{i=1}^{n}x_i^2-\frac{1}{n}(\sum_{i=1}^{n}x_i)^2} \qquad a=\frac{1}{n}\left(\sum_{i=1}^{n}y_i-b\sum_{i=1}^{n}x_i\right)$$

In any case number of parameters cannot be greater than number of x_i,y_i pairs i.e. $k \ n$ [16].

The task here is presented as mathematical analysis task assuming that all quantities (variables) have no uncertainty. But it can be extended: If uncertainties (variances, covariances) are included it is a statistical problem of propagation of uncertainties and uncertainties of parameters can be gained.

Regression is used for many tasks like verification of theoretical formulas. So called smoothing used to reduce scatter of data is a special frequent case. The number of points entering the regression process is reduced to *n = 3-5* on "both sides" around a missing point and $P^{a,b}(x)=ax+b$ approximation function is linear.

[16] In the opposite case it is so called underdetermined task and additional information should be added.