Chapter 14
Learning from Residuals

Residuals often are considered as a troublesome noise in spatial—or, for that matter—non-spatial econometric models. Current practice in spatial econometrics is to set up a spatial error model, more often than not with an exogenous W spatial weight matrix, in order to improve the efficiency of the estimators.

Looking closely into the residuals is less common practice. And still, residuals can represent extremely precious building blocks for further work, as other disciplines have shown. Around 1850 the British chemists, Mansfield and Perkin, had the—for that era of chemistry—strange idea to analyze the composition of tar, until then exclusively used to improve coverage of roads (John London McAdam had his name attached to that technique, tarmacadam); the result of the British chemists' investigation was the roaring development of a whole branch of (industrial) chemistry: carbochemistry.

In the next section, a simple spatial econometric example will be treated, after which further analysis and more results will be presented.

14.1 Residuals

Tables 14.1 and 14.2 present the degrees of contiguity for Belgian regional units, BRU, (the maximum degree being 3) and their gross regional products (1995, 10^5 Euros of 2000); the entries of the two tables follow the same order.

Figure 14.1 reproduces the map of those regions.

The regions are the following. From West to East, northern slice: West-Flanders, East-Flanders, Antwerp, Limburg; same, southern slice: Hainaut, Namur, Luxembourg, Liège (slightly upwards); right in the middle, from north to south: Flemish Brabant and Walloon Brabant, with the Brussels Capital region sticking out.

First the products of 1995 were analyzed. The idea was to investigate the effects of (average) products for different degrees of contiguity (1, 2, 3) on a given GRP, y_i. Hence the equation

$$y_i = ay_{1i} + by_{2i} + cy_{3i} + d + \varepsilon_i, \tag{14.1}$$

Table 14.1   Degrees of contiguity between Belgian regions

BRU	A	BW	VB	OV	WV	LIM	H	N	LU	LIE	BC
A	0	2	1	1	2	1	2	3	3	2	2
BW	2	0	1	2	2	2	1	1	2	1	2
VB	1	1	0	1	2	1	1	2	2	1	1
OV	1	2	1	0	1	2	1	2	3	2	2
WV	2	2	2	1	0	3	1	2	3	3	3
LIM	1	2	1	2	3	0	2	2	2	1	2
H	2	1	1	1	1	2	0	1	2	2	2
N	3	1	2	2	2	2	1	0	1	1	3
LU	3	2	2	3	3	2	2	1	0	1	3
LIE	2	1	1	2	3	1	2	1	1	0	2
BC	2	2	1	2	3	2	2	3	3	2	0

Same obvious of province abbreviations given also in Table 14.2

Table 14.2   Gross regional products for the Belgian units, 1995

Regional units	A	BW	VB	OV	WV	LIM	H	N	LU	LIE	BC
Values	416028	62919	211584	255118	226222	143460	191433	64851	37976	173063	424381

Fig. 14.1   Regional map of Belgium

Table 14.3 First results for model (14.1)

Parameters	Values	t- or F-values	Probability
a	1.8813	1.8900	0.1007
b	1.4596	1.4349	0.1945
c	–0.0378	–0,1198	0.9080
d	–411581	–1.0873	0.3129
R^2	0.4806	2.1590	0.1810

where y_{1i}, y_{2i} and y_{3i} are the average products for different degrees of contiguity.

Table 14.3 presents the OLS estimation results.

Obviously the results are far from being satisfactory. The residual spatial correlation coefficients, r_c^2 ($c=1, 2, 3$, the observed degrees of contiguity) are respectively *–0.2619, –0.1161* and *–0.2426*. They are not significant, but show that there is no completely random field in the residuals.

Accordingly, further analysis is in order.

14.2 Multiple Regimes

The first column of Table 14.4 shows the residuals of the exercise, and compares them (columns 2 and 3) with the growth rates (averages over 1995–2004) and the GRP levels.

The following scatter plots (Figs. 14.2 and 14.3) picture the partial relations.

The *Kendall-τ* (Kendall, 1955) between residuals (+ or –) and growth rates (above or below the average, 0.0217) is near zero (exactly, 0.0910), and between residuals and GRPs it is 0.4546, but further investigation is still required.

To prepare the latter, a complexity index has been computed (Getis and Paelinck, 2004), derived from a fourth degree polynomial

Table 14.4 Comparing residuals

| Regional units	Residuals	Growth rates	GRP
A	132145	0.0202	416028
BW	−159733	0.0335	62919
VB	15140	0.0277	211584
OV	14406	0.0239	255118
WV	−50515	0.0207	226222
LIM	−103617	0.0194	143460
H	−54674	0.0144	191433
N	−31766	0.0245	64951
LU	15883	0.0190	37976
LIE	82491	0.0135	173063
BC	140241	0.0221	424381
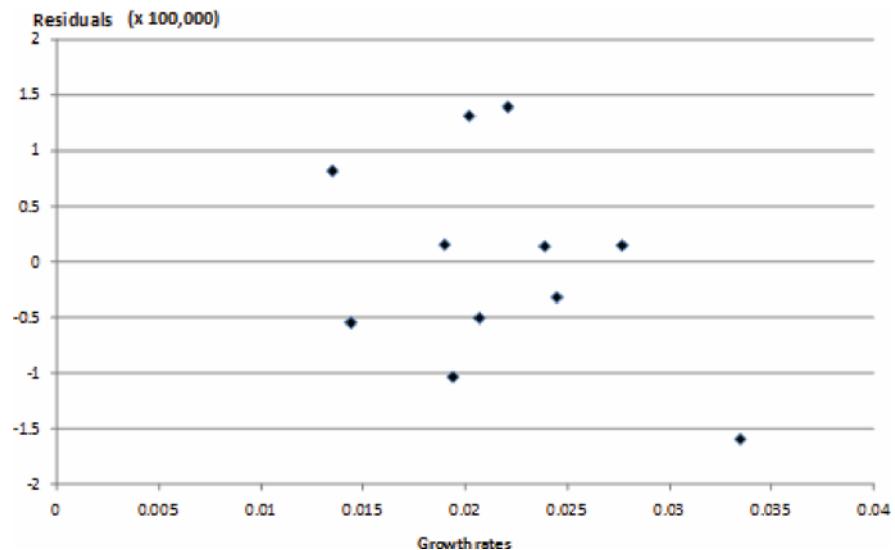
Fig. 14.2 Residuals and growth rates from Table 14.4
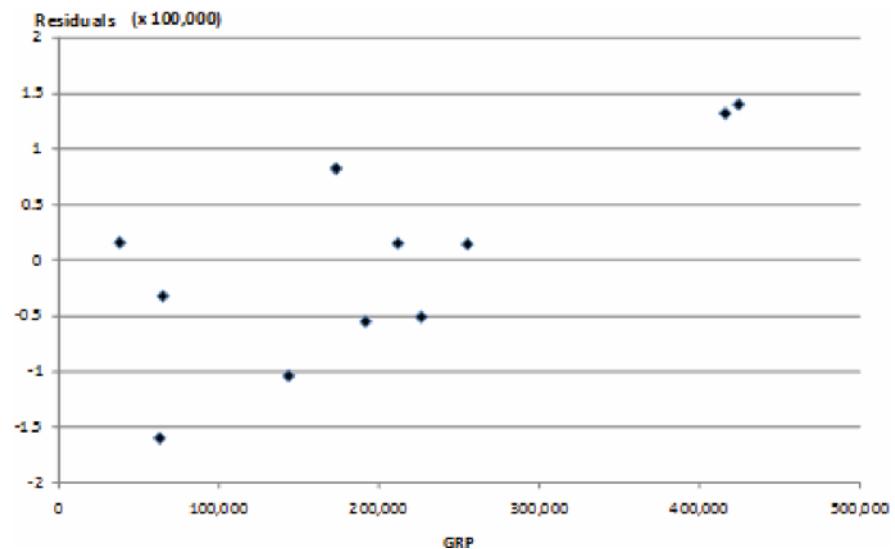
Fig. 14.3 Residuals and GRP from Table 14.4

Table 14.5 Polynomial coefficients from Eq. (14.2)

Coeffic-ients	a	b	c	d	e	f	g	h	i	j	k
Values	1740457	–2.7E+08	55.7706	1.54E+10	–0.00025	–5903.13	–3.01E+11	–1.45E-11	148311	0.02579	–0.62572

$$e_i = a + b*r_i + c*y_i + d*r_i^2 + e*y_i^2 + f*r_iy_i + g*r_i^3 + h*y_i^3 + i*r_i^2y_i + j*r_iy_i^2 \\ + k*r_i^2y_i^2,$$

(14.2)

in which the y_is are again the GRPs, and the r_is the growth rates. Table 14.5 presents the interpolated coefficients of Eq. (14.2).

Coefficients *b, d,* and *g* are extremely high, but they relate to growth rates that are small numbers. Excluding the relatively small coefficients (smaller than one), the complexity coefficient can be computed as

$$C = (v - 1)/(n - 1) = 0.6,$$

(14.3)

where v is the number of maintained coefficients, and n their maximal number (i.e., the number of observations).

Given the rule followed, this is a relatively high value $(0 \leq c \leq 1)$, and invites rethinking the model generating the observed residuals, as this model is a very simple one.

The revealed complexity suggests the need for a possible correction by r_i and y_i, but the first correction would not be complete, as shown above, and using y_i would be trivial. A plausible alternative would be to introduce two separate regimes, leading to the following specification (see Chap. 12):

$$y_i = \lambda_i(a*y_{1i} + b*y_{2i} + c*y_{3i} + d) + (1 - \lambda_i)(\alpha*y_{1i} + \beta*y_{2i} + \gamma*y_{3i} + \delta) + \varepsilon_i, \quad (14.4)$$

where the λ_is are the binary variables qualifying the spatial regimes.

This produced the results of Table 14.6 hereafter.

Table 14.6 Results with two regimes, Eq. (14.4)

Parameters	Values 1995	Values 2004
a	0.6726	0.5894
b	1.6484	1.5462
c	–0.7094	–0.7416
d	–69490	–763
α	8.5571	2.9617
β	4.7365	1.0460
γ	–0.1971	1.2467
δ	–2345031	–710212
R^2	0.9853[a]	0.9774[a]

[a]Computed for 6 df

Table 14.7 Regime
allocators for two distant
years

Regional units	1995	2004
A	1	1
BW	0	0
VB	0	0
OV	1	1
WV	1	1
LI	1	1
H	0	0
N	1	1
LU	1	0
LIE	1	1
BC	0	1

The residual spatial correlation coefficients for contiguities 1, 2 and 3, respectively, are *–0.1793, –0.8029* and 0.2938, showing that there is still some specific spatial autocorrelation, especially of order 2 (coefficient significant at the 0.995 level). Some changes occurred over nine years, as the third column of Table 14.6 shows; this also was the case for the r_cs (–0.6702, –0.8223, 0.6428, all significant at the 0.995 level), the second order spatial autocorrelation still dominating. But the overall fit is satisfactory, and OLS can be replaced by other estimation methods (see Sect. 11.1.3).

To test the general properties of the residual fields, two statistics have been computed:

– a generalized τ-statistic between all residuals (i.e., 55 cross-products are involved); for 1995 and 2004 they amount to non-significant values –0.1542 and –0.1523, respectively, excluding any general correlation; and,
– the C-statistics of Eq. (14.3); in both cases they equal 0.9, showing a high degree of Chaitin-Wolfram complexity (as independent variables, the numbers from 1 through 12 were used in the test polynomial).

The problem now is: though overall randomness seems to be present, spatial r_cs show specific dependency, so some further investigation is in order.

Table 14.7 presents the two vectors of λ_I estimates.

The pattern is remarkably stable: most regions (7) belong to the same regime, only the center-south deviating from this.

14.3 Spatial Interpolation

Because there is quite some variability in the coefficients reported in Table 14.6, the question arises asking whether computing local coefficients could give more insight in this phenomenon. One possibility is to interpolate the parameters from

Table 14.8 Results from spatial interpolation to compute the parameters of Eq. (14.1)

Parameter regional	a	b	c	d	a	b	c	d
A	2.8428	3.2064	−0.0712	816849	2.9034	3.0462	−0.8971	983912
BW	1.1572	−0.4494	0.0776	0	1.0961	−0.4499	0.0747	0
VB	9.0711	5.0851	−0.3824	2.5058	8.2277	4.6305	−1.2307	2.6980
OV	3.5057	3.6279	−1.1336	1.0209	3.3062	3.2851	−1.1589	1.1109
WV	2.1252	1.0591	−0.1991	410515	1.6838	0.5410	−0.0943	282235
LI	1.5086	2.6832	−0.5342	441860	1.2176	2.2869	−0.6389	380635
H	8.5517	4.7330	1.0080	−2.3433	6.0211	2,8355	0.7076	−1.8282
N	6.0664	0.1237	0.3975	−833866	4.0873	0.1847	0.2981	−673717
LU	−1.6767	1.0559	−0.4168	214303	−1.7254	1.1119	−0.4436	254356
Lie	.6890	1.3824	−0.5545	71376	0.6151	1.2632	−0.4947	76625
BC	9.0717	5.0851	−0,3822	2.5058	8.2277	4.6305	−1.2307	2.8980
v=σ/μ	0.9106	0.7594	−2.6792	1.5842	0.9346	0.7818	−1.3398	1.7597

groups of— possibly neighboring—spatial units. If more than four regions present themselves as candidates, nearest neighbors—in terms of distances and/or political/linguistic proximity—have been selected.

Table 14.8 presents the results. The parameters are those of Eq. (14.1).

The remarkable finding, again, is that the orders of magnitude are the same for the two years, with some exceptions for the constant *d.* But the variability is large between regions (as shown by the coefficients of variation in the last row of Table 14.8), which suggests the need for further analysis of the available data to complete the picture.

14.4 Composite Parameters

Because time series from 1995 through 2004 are available, composite parameters can be computed (Ancot et al., 1978); for instance, parameter a in Eq. (14.1) can be expanded as

$$a = \hat{a} + a_r + a_t, \tag{14.5}$$

Table 14.9 Generic and region-specific coefficients according to model (14.5)

Region Parameter	A	BW	VB	OV	WV	Lim	H	N	Lu	Liè	BC
$a(a_r)$	– 1.26	1.113	0.6643	1.715	1.238	1.249	–0.8	2.055	– .0486	– 0.230	1.298
$b(b_r)$	3.132	– 2.79	0.1305	– 2.39	– 2.39	– 2.88	– 0.93	– 3.27	– 3.59	– 2.40	– 1.92
$c(c_r)$	– 0.065	–	–	– 67.8	0.5279	0.4333	–	– 0.493	1.086	0.4754	2.51
$d(d_r)$	132	–97	– 116	–67	6.342	– 212	41	7.877	–22[a]	0.0326	– 0.337

[a]To be multiplied by 10^3

where a is the generic, a_r is the region-specific parameter, and a_t is the time-specific parameter. For reasons of identifiability, one spatial unit should be selected as a kernel (not affected by a_r or a_t); the first region, Antwerp, was picked for this purpose, but any other region would have done.

Table 14.9 hereafter presents the coefficients; the first row, as said above, contains the generic ones, the following rows the region-specific ones. Sometimes a region-specific coefficient c_r is absent, due to the absence of a third order spatial lag.

The entries of Table 14.9 display a large region-specific parameter variability. No measure hereof has been computed this time, but a comparison with Table 14.8 confirms this variability.

Table 14.10 presents the time-specific parameter estimates for 1996 through 2004.

The coefficients are of a much smaller order of magnitude, which confirms a previous remark about the relative constancy of the parameters through time, as opposed to their interregional variability.

Table 14.10 Time-specific parameters for model (14.5)

Parameter FUTL[a]	a_t	b_t	c_t
1	−0.0026	0.0094	−0.0153
2	−0.0016	0.0053	−0.0016
3	−0.0097	0.0206	0.0028
4	0.0044	0.0188	−0.0205
5	0.0050	0.0209	−0.0317
6	0.0028	0.0304	−0.0365
7	0.0122	0.0220	−0.0489
8	0.0153	0.0218	−0.0513
9	0.0176	0.0199	−0.0588

[a]Forward Unit Time Lag, from 1995 on

Table 14.11   Pseudo R^2s

Regions	Pseudo-R^2s
A	0.9538
BW	0.9905
VB	0.9907
OV	0.9650
WV	0.9993
LI	0.9926
H	0.8705
N	0.9861
LU	0.9394
LIE	0.9104
BC	0.9809
Global	0.9735

Finally Table 14.11 presents the partial and global pseudo-R^2-values, pseudo-because the parameters have been computed by least absolute discrepancies to avoid outliers.

The result is remarkably high for 60 df, with a local exception for Hainaut.

14.5 Conclusion

The doggy-bag principle ("never throw away your leftovers") has given insight into a possibly appropriate specification of the spatial econometric models investigated. This is in line with the clear warning that has been given off for time series analysis (G. Mizon, A Note to Autocorrelation Correctors: Don't, Journal of Econometrics, 1995, 69, pp. 267–288).

More research is in order, especially for very large models. But considering residuals as informative should transcend the usual practice of trying to neutralize them. Meanwhile, pure spatial "randomness" also could be interpreted as spatial complexity, and might encourage continued analysis rather than finishing it by discussing "ideal" parameter properties.

In the Belgian case, this has lead to deeper insights in spatio-temporal properties of a static model. Indeed, it appears that each spatial unit possesses its own reaction coefficients with a great stability over time. Problem however is to find out how much of that interregional divergence is due to system heterogeneity, and how much to spatial aggregation. The latter problem is taken up in Chap. 17.