Quantifying Social Fields by Turning Regression Modeling "Inside Out"

Ronald Breiger, University of Arizona

to be presented at the Quantifying Social Fields Conference,

University of California, Berkeley, April 20-21

April 11, 2012

Dear fellow participants in the Quantifying Social Fields Conference,

I am pleased to provide the attached paper (presented just last week, on April 4, at the International Studies Association annual meeting) as "read-ahead" material for the paper I will be presenting in Berkeley on April 20-21 on behalf of my research group. Here is what we think we are doing that is new.

We are developing a "dual" to regression analysis, one in which the key results of the regression (including the partial regression coefficients and the predicted values on Y) may be seen to depend on the *cases*. This allows us to use the variables to learn about the cases, whereas the standard use of regression analysis is to make the cases invisible, and to discuss only the relations among the variables (see also Shalev 2007 for a critique of the use of regression analysis in the comparative study of welfare states). Unlike many critics of "general linear reality" (Abbot 1988, and, subsequently, many other critics), we do not want to overturn regression modeling. To the contrary, we want to get more out of it. We do so by recognizing foundations of regression modeling in "spaces" similar to those used by practitioners of case-oriented techniques (Ragin, 2008), multidimensional scaling (Shalev, 2007), field theory, and correspondence analysis (Le Roux and Rounet, 2004).

Research on network modeling, and insights from sociological field theory, may be applied to the network among the cases, and to a dual cases-variables network that, as we show, underlies the usual regression modeling. Doing so leads to new discoveries about the organizational and relational underpinnings of regression models and their applications. Your comments are most welcome.

Sincerely,

Ron Breiger

Application of a Profile Similarity Methodology to Leverage Open Source Data on CBRN Activities of Terrorist Groups

Ronald L. Breiger

David Melamed

Eric Schoon

University of Arizona

Paper presented at the Annual Convention of the

International Studies Association

San Diego, April 4, 2012

This work was supported by the Defense Threat Reduction Agency, Basic Research Award # HDTRA1-10-1-0017, to the University of Arizona. Eric Schoon was supported by a National Science Foundation Graduate Research Fellowship. We thank Lauren Pinson of the START Center for consulting with us about the POICN database.

Application of a Profile Similarity Methodology to Leverage Open Source Data on CBRN Activities of Terrorist Groups

Ronald L. Breiger, David Melamed, and Eric Schoon

University of Arizona

1. From Information to Analysis

Important studies of terrorist social network connections have been conducted on full network data ("who-to-whom" and "who-to-what") derived from open sources (e.g., Krebs, 2001; Pedahzur and Perliger, 2009; Roberts and Everton, 2011; Rodriguez, 2005; at the level of states, see Asal et al., 2012). Nonetheless, in many situations information on the ties among terrorists is notoriously "incomplete, inaccurate or simply not available" (Tsvetovat and Carley, 2005; see also Hayden, 2009; Sparrow, 1991). One highly productive reaction has been to focus on computational modeling in order to understand behavior on the basis of simulated terrorist networks (e.g., Tsvetovat and Carley, 2005). In this paper we pursue a different strategy, making use of database information on actual groups and some of their known behaviors and attributes.

As Perliger and Pedahzur (2011) point out, there has been "a striking increase in efforts and resources invested in data collection" on terrorist groups in recent years by academic and governmental agencies. Particularly notable in this respect have been the open-source, publicly available databases maintained at the START Center at the University of Maryland, resulting in the present availability of "high-resolution" information (see also Hayden, 2009).

Among the newest databases produced by the START Center is POICN, the Profiles of Incidents involving CBRN by Non-state actors. Described in greater depth in a paper presented at this panel (Ackerman and Pinson, 2012), the START researchers began with 499 potential cases drawn from currently existing databases. They created 142 core variables relating to the geo-spatial, temporal, motivational, operational, and tactical consequence aspects of each CBRN incident. (CBRN pertains to chemical, biological, radiological, or nuclear weapons.) Of particular interest to researchers: Most of the core variables in the POICN database feature measures for validity, thus allowing researchers to explore the effects on their model coefficients of using varying standards for how credible an event description needs to be in order to be included.

We make use of 181 incidents in the POICN database related to CBRN attempted attacks and successful attacks. We show how a form of modeling often used by social network analysts can be applied to leverage information in distinctly new ways from databases on terrorist activities. Specifically, we derive network connections from profile similarities among incidents in this database. In the process we shed new light on the conventional regression modeling that has been a main mode of analysis in political science, sociology, and throughout international studies.

2. Context and Research Questions

Given the absence to date of a true mass-casualty WMD attack by terrorists, what can be gained be examining empirical evidence from the broader category of CBRN weapons use? We would put forward two reasons for an interest in the types of attack detailed in the POICN database. First, uses and attempted uses even of minor quantities of CBRN materials or in plots resulting in no deaths and in minor damage can still provide an indication of planning and / or evolving capabilities that could lead to more consequential operations. Second, small-scale use can blur the line between CBRN and conventional weapons use with respect to what is acceptable to perpetrators (cf. Schelling's classic analysis, 1960, on a different plane: maneuvering among Cold War nation states over definitions of what constitutes a WMD activity).

The outcome variable upon which we focus is the total number of all individuals who died as a direct result of a CBRN event. Of the seven CBRN event types coded in the database (ranging from proto-plots, plots, and attempted acquisition all the way up to use of an agent), the number of deaths is coded only for events involving the actual or attempted use of an agent. The number of events involving actual or attempted CBRN use was 186 (reduced to 181 after cases exhibiting missing data on the predictor variables were excluded). The vast majority of these events (about three-fourths of them) resulted in no deaths; 43 events resulted in from 1 to 200 deaths each.

In our frankly exploratory analysis we made use of a range of predictor variables drawn from the database that might reasonably be related to the total number killed in an event. We included type of agent (whether chemical, biological, or other), other attack characteristics (whether each event was primarily an assassination attempt, aimed at holding hostages to gain political concessions, featured use of explosives), and characteristics of perpetrators (whether some level of the CBRN agent was produced "in house," and whether the perpetrators included a cult, a religious extremist group, a one-issue group such as environmentalism, a lone individual or cell not linked operationally to other groups, or an ethno-nationalist group, or some other type of group). Table 1 provides brief descriptions of these variables as well as means (and standard deviations) or proportions.

TABLE 1 ABOUT HERE

In a conventional multivariate study we would pose our central research question as follows:

> **RQ1**: Which variables (which combinations of attack characteristics and perpetrator characteristics) best predict the lethality of attacks?

Indeed we are interested in that question. In addition, however, our research questions move beyond a focus on how variables relate to other variables. We are also interested in how we can use the variables to learn about the cases (the CBRN events). Thus we also formulate

> **RQ2**: Identify clusters of cases that have distinctive patterns of contribution to the regression coefficients of the standard regression model. What *are* those distinctive patterns, and what can we learn from the *multiplicity* of these patterns about "different CBRN stories" (i.e., distinctively different patterns of regression coefficients within the overall regression model)?

3. Analysis: Conventional Regression

As mentioned above, about three-fourths of our 181 events exhibit zero casualties. The appropriate model to use with these data is a zero-inflated Poisson model (Gelman and Hill, 2007, pp. 126-27). However, in order to focus this presentation of results on the innovations of our approach, we illustrate an application of ordinary multiple regression (OLS) after having transformed all variables to standard form (mean of 0, standard deviation of 1). Although it is not a realistic model for these data, at least OLS is fairly robust. We will demonstrate how we can gain insight even from application of the most basic regression model.

Linear regression coefficients are given in Table 2. The adjusted R-square is a fairly modest .10 across the 181 cases. Results indicate that events whose perpetrators are cult members have 26.8 more deaths on average (standardized coefficient = .385, p = .002), net of the other variables in the model. Hostage-taking also leads to increased deaths (72.4 additional deaths; standardized coefficient = .220, p = .003). Events in which a CBRN agent was produced “in house” have on average 18.2 *fewer* deaths, net of the other variables (standardized coefficient = -.275, p = .02), and events in which assassination is the primary aim have on average 12.2 fewer deaths (standardized coefficient = -.172, p = .05). Events that the POICN coders doubted were true cases of terrorism had on average 6.6 *more* deaths, net of the other variables (standardized coefficient = .133, p = .09). Events involving explosives had 7.6 more deaths, net of other variables (standardized coefficient = .117), though this was not significant (p = .21).

TABLE 2 ABOUT HERE

The portrait drawn by these results suggests that the production of CBRN agents “in-house,” and also the goal of assassination, are associated with fewer deaths. In the latter case, assassination (even when it is realized) typically kills one or a very small number of people, and in the former case, perhaps it is the smaller and less well-resourced groups that produce CBRN agents very locally. Conversely, hostage-taking and participation by cults, both of which lead to higher deaths, are both features of groups considerably more disciplined. The net association of cases the coders doubted to be terrorism with higher deaths per event (not quite a significant finding, p = .09) suggests that there is a consequential class of CBRN use that is different from terrorism.

Without in any way contradicting these results and findings, we now present our approach to turning the usual regression modeling “inside out” in order to focus on how these results concerning *variables* may be seen to depend on networks among the *cases*, which here are our 181 CBRN events. In contrast to the average effects given by the regression coefficients (reviewed above and in Table 2), we will identify multiple clumps of cases, each refracting the overall regression results in a distinctive way. This will lead us to a more nuanced interpretation of the effects of our predictors on the lethality of the events.

4. Analysis: Turning the Regression “Inside Out”

a) Motivation. Given a data matrix (cases by variables), regression analysis as well as many of its generalizations may be thought of as the study of relations among the *variables*. With its typical assumption that the “cases” are a random sample representative of a population of interest, regression analysis makes the cases invisible, as Michael Shalev (2007) and other analysts of comparative politics have argued in their critiques of regression approaches.

But often the cases *are* of interest, and the goal of the analysis should be to use the variables to let the cases be seen. Shalev (2007) discusses analyses where the cases are countries, and the research agenda is comparative analysis of types of welfare states. In the example of the present paper, the cases are CBRN events, and our research agenda is comparative analysis of types of such events (discovering the types and how variables interact differently within each type). Moreover, in neither Shalev’s examples nor those of the present paper could the analyst claim that the cases are a random sample. The POICN database aims at collecting *all* known cases of CBRN events within its date range, and there are surely dependencies among the events along multiple dimensions. (For example, two attacks attempted by the same group in adjacent months are likely not “independent” of each other. In addition, attacks using the same toxic agent by different groups but within the same country might well lack independence from one another. It seems quite limiting indeed to assume independence among all the cases.) We propose instead to discover regions of *dependence* among the cases on the basis of their attributes. Along with Shalev (2007), Charles Ragin (2008), and other researchers, we are willing to pay the costs of giving up our claim to “significance testing” in order to gain insight by more richly exploring the structuring internal to our dataset. Moreover, we show that we can do all this by deepening the framework within which regression analysis is conventionally understood.

b) Some formal shorthand. Consider an $n \times p$ data matrix (denoted **X**) whose rows represent each of the n cases (in our example, 181 CBRN events) and whose columns stand for the p predictor variables (p = 12 variables in standard form in our example). Assuming a continuous outcome variable, **y** (n × 1), matrix notation for the fitted values of **y** (denoted $\hat{\mathbf{y}}$) in the linear regression model is

$$\hat{\mathbf{y}} = \mathbf{X}\,\mathbf{b} \qquad (1)$$

where **b** is a $p \times 1$ vector of regression coefficients estimated by the ordinary least-squares criterion. We compute the singular value decomposition (SVD) of **X**,

$$\mathrm{X} = \mathbf{U}\,\mathbf{S}\,\mathbf{V}^{\mathbf{T}} \qquad (2)$$

the point of which (for our purposes) is to produce (as the columns of **U**) a set of orthogonal dimensions pertaining to the rows of **X** (the terrorist organizations), and (as columns of **V**) a set of orthogonal dimensions for the columns of **X** (the predictor variables). (Superscript **T** denotes matrix transposition.) **S** is a diagonal matrix of weights (singular values) indicating relative importance of each dimension. The regression coefficient, b, are often estimated as

$$\mathbf{b} = (\mathbf{X}^{\mathbf{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathbf{T}}\mathbf{y} \qquad (3)$$

but by substituting (2) into (3) we see that the identical regression coefficients may be written

$$\mathbf{b} = \mathbf{V}\ \mathbf{S}^{-1}\mathbf{U}^{\mathrm{T}}\mathrm{Y}\ \mathbf{1} \qquad (4)$$

where **Y** = diag(**y**) is a diagonal matrix of observed values on the outcome variable, and **1** is a vector of 1's.

Equation (4) is interesting to us for two main reasons (Melamed et al., 2012). First, $\mathbf{V}\ \mathbf{S}^{-1}\mathbf{U}^{\mathrm{T}}\mathbf{Y}$ is of size variables by cases (p x n), and each of its rows reveals the contribution of each *case* to a given *regression coefficient*. We are unaware of other researchers who exploit this fact, and we will do so in discussion of our Table 3 below.

Our second source of interest in eq. (4) is, analogously to researchers who employ principal components regression (Gunst and Lee, 1980), the vector **b** of regression coefficients may be seen to be composed from a series of dimensions:

$$\mathbf{b}_{\mathrm{k}} = \mathbf{V}[,1:\mathrm{k}]\ \mathbf{S}[1:\mathrm{k},1:\mathrm{k}]^{-1}\mathbf{U}[,1:\mathrm{k}]^{\mathrm{T}}\mathbf{Y}\ \mathbf{1} \qquad (5)$$

When $k = p$ (the number of variables), equation 5 is identical to the previous one. But when $1 \leq k < p$, we have (by the least-squares principle, and in the sense of principal components regression) a "best" representation of the regression coefficients in the first k dimensions.

Going along with the above, we can see that $\hat{\boldsymbol{y}}$, the fitted values of our outcome variable, are composed from a series of dimensions:

$$\mathbf{U[,1:k]\ t(U[,1:k])\ y} = \hat{\boldsymbol{y}}_{\mathrm{k}} \qquad (6)$$

When $k = p$, eq. (6) becomes

$$\mathbf{[U\ U^{\mathrm{T}}]\ y} = \hat{\boldsymbol{y}} \qquad (7)$$

The expression in brackets is a network among the cases (of size n x n) based on profile similarity (Breiger et al, 2011). Eq. (7) “says” that the predicted value from ordinary regression for (let us say) the first-listed case is identical to the sum of the observed value on Y of each case, multiplied by each (respective) case’s similarity to the first. Eq. (6) tells us that the predicted values on Y may be decomposed into best-fitting (in the sense of least-squares) components based on using the first *k* dimensions.

The correlation of $\hat{\boldsymbol{y}}$ with $\hat{\boldsymbol{y}}_k$ (for k = 1, …, 12) is reported in Figure 1. The 12 correlations reported there correlate -.94 with the diagonal of the **S** matrix (see eq. 2), suggesting the role of dimensions in decomposing the fitted Y-hat values. From Fig. 1 we see that using just the first dimension of the data produces $\hat{\boldsymbol{y}}_k$ values that correlate .18 with $\hat{\boldsymbol{y}}$**,** whereas using the first two dimensions produces a set of $\hat{\boldsymbol{y}}_k$ that correlate .36 with $\hat{\boldsymbol{y}}$**.** In the following, we will often use a two-dimensional representation of the regression model due to the simplicity of the two-dimensional representation, but we will be mindful of how much of the regression model we are thereby leaving aside (i.e., the correlation of .36 is quite low; see Fig. 1.)

FIGURE 1 ABOUT HERE

c) The regression model in 2 dimensions. We may adjoin $\hat{\boldsymbol{y}}_2$ (eq. 6) to the data matrix, **X**, and then compute the SVD on [X | $\hat{\boldsymbol{y}}$] (of size *n* by *p* + 1). We use the notation **U*** to symbolize the resulting row space (compare eq. 2, the SVD on **X** alone). Each of the first *p* columns of **U*** is correlated perfectly with the corresponding column of **U**. A graph of **U*** and of **V** is given in Figure 2. Also shown there (labeled “Yhat2dim”) is $\hat{\boldsymbol{y}}_2$ (eq. 6), as well as projections of each variable to the line connecting $\hat{\boldsymbol{y}}_2$ to the origin of the graph (point (0,0)).

FIGURE 2 ABOUT HERE

One implication of Figure 2 is that, from it, we can read ratios of the regression coefficients (or, more precisely: their two-dimensional representations, $\hat{\boldsymbol{y}}_2$)**.** For example, these two-dimensional regression coefficients for the variables CHEM (use of a chemical agent) and "ethNat" (a perpetrator from a group coded as ethno-nationalist) are .0626 and -.0073 (respectively). We see that, in this two-dimensional representation of the model, use of a chemical agent has 8.0971 times the impact on total deaths as does an attack involving an assassination attempt (.0626 / -.0073 = -8.0971; the signs imply that use of a chemical agent increases the casualty rate, while having a perpetrator from an ethno-nationalist group slightly decreases it). Looking in Fig. 2 at the distances of the projections of these variables from "Yhat2dim" gives exactly the same ratio. Specifically: the ratio of distances is (.5840 / .0721) = 8.0971 also. Thus, pictures such as Figure 2 allow us to visualize the relative effects of regression coefficients on the dependent variable (given the dimensionality of the figure, in this case 2 dimensions). In this sense (and in others), pictures such as Figure 2 are "natural" adjuncts to standard regression modeling.

So what? The point of all this is that we see from Figure 2 that some clumps of variables "hang together" with respect to their net effects on Y, such as religious extremist groups and using explosives. Another clump that hangs together would seem to include cult groups and producing CBRN agents "in-house." Moreover, these two sets of variables are far apart from each other. This suggests that clustering of the variables actually underlies the usual regression model, even though such thinking has (we believe) never been brought to bear in analyzing regression equations. And, the same applies to cases.

d) Cases in different regions of the regression space. A main motivation of our approach is to bring cases (CBRN events) and variables into the same "picture" of the regression model space. To do so is to recognize (deductively) and to discover (inductively) that the regression space is often inhomogeneous. Consider for example POICN database case # 517. The POICN database narrative of this case reads as follows:

> In late November [1994], Aum Shinrikyo cult members Satoru Hirata, Yoshihiro Inoue Tomomitsu Niimi and Akira Yamagata attempted to assassinate Noboru Mizuno at his home in Nakano Ward, Tokyo, Japan (A, B). Mizuno was helping deserters from Aum Shinrikyo escape and reportedly also tried to sue Aum, which made the cult want to get rid of him. Hirata tried to spray VX nerve agent, from a syringe, on Hirata's neck. The attack was unsuccessful due to reasons most likely related to how the agent was made (A, B). Aum produced VX nerve agent at one of their chemical laboratory facilities. Shoko Asahara, cult guru, ordered the hit. The group made a second attempt on Mizuno's life, also with VX, the following month (A, B). The assassination attempt may not always be considered terrorism but rather seen as an assassination attempt to remove an obstacle that may have prevented Aum from achieving its goals. [Source: POICN database. Letters in parentheses such as (A,B) refer to sources given in the database record for this case.]

The POICN database coders saw in this narrative certain variables of interest: a cult group; a chemical agent; the production of that agent "in-house"; an attack that was primarily an assassination attempt; and doubt that the event was an act of terrorism. Please see Fig. 3, where the coordinates of this event (from matrix U, eq. 2 above) are given, along with connections of this event to the coded properties (from matrix V of eq. 2).

FIGURE 3 ABOUT HERE

Let us contrast the previous event with one other: POICN database case # 280. The narrative here is as follows:

> On July 15, 2003, a homemade explosive device was found placed on tanks of acetone and phenol at the Saratovorgsintez Limited Liability Company in Saratov, Russia (A, B, C, D). The device was made from a modified RGD-5 hand grenade, with extra cartridges and wiring. The device was deemed harmless after police removed it from the chemical company and discovered it was made with a training grenade and was not a live explosive (A, B, C, D). Ethnonational Chechen Rebels were suspected to have placed the device at the chemical plant to intimidate Moscow (A). Some reports indicate that authorities believed the Chechens were planning bombings in at least four other cities in Russia (A).

Although this case too involves chemical agents (CHEM, referring to the tanks of acetone and phenol on which the explosive was placed), this narrative “feels” like a different case than the previous one. Here the coders found, in addition to CHEM, the use of explosives by an ethno-nationalist group. Indeed, this case occupies a different region of the cases-by-variables “space,” as indicated in Figure 4.

FIGURE 4 ABOUT HERE

Examination of specific case narratives such as these two encourages us to look for a clustering of all 181 cases based on their attributes. Toward this end we apply a standard clustering procedure (k-means) to the **U** matrix of eq. 2 (see Melamed, Breiger, and Schoon, 2012, for a discussion of this point). The procedure clearly discovers distinctive clumps of cases, as illustrated in Figure 5.

FIGURE 5 ABOUT HERE

Somewhat arbitrarily, we chose a six-cluster solution. We partitioned the columns of $\mathbf{V}\ \mathbf{S}^{-1}\mathbf{U}^{T}\mathbf{Y}$ (see eq. 4 and related discussion), which is a matrix of 12 rows (one for each variable) and 181 columns (one for each event), into a 12 x 6 (clusters) matrix. Table 3 shows the results. Notice that Table 3 illustrates a key feature of our approach, namely that

> **the usual regression coefficients (in Table 2) are sums across clusters of cases!**

TABLE 3 ABOUT HERE

Thus, contrary to the way we learned about regression in our mandatory “stats” class, the regression coefficients may be understood as relating (clumps of) cases to variables.

Let’s now take a look at Table 3.

5. What we learn about CBRN events from this analysis

Beginning with Cluster A in Table 3, we see that about half the events in our study (96 events) formed a cluster that contributed to the overall regression coefficients in a way that is (with a few exceptions) consistent with the estimated coefficients from the standard regression in Table 2. (For this first cluster, for example, the net effect on lethality of assassination and of in-house production was negative; the net effect of perpetrators from cults was positive; and so forth).

However, for the other half of the events in our study, the interpretations diverge from that of the standard regression coefficients in Table 2. With respect to hostages, virtually the entire net effect of “hostages” on “total deaths” (the regression coefficient in Table 2 being .2197) is due to Cluster B (see row 2, column 2 of Table 3), and Cluster B consists of a single case (case # 174, an event involving the FARC).

Usually we would recognize a single deviant case (the single case in Cluster B) as an outlier, and we have a variety of well-known techniques for dealing with outliers (Belsley, Kuh, and Welsch, 1980). However, it is our view that (often in general) an entire set of datapoints consists of sets of partially overlapping outliers.

In this application, for example, we find that almost the entire net effect of explosives on number of deaths comes, not from Cluster A, but from Cluster F (see the “expl” row of Table 3). Cluster F is composed of 42 events, with an overrepresentation of events having ethno-nationalist groups as perpetrators. (To take one example, event # 280, narrated above, involving allegations of Chechen rebels placing an explosive on top of cans of chemicals, is placed in Cluster F by our clustering procedure.) Cluster F seems to be a distinctive “clump” of cases in which there is a very high net effect of explosives use on casualties, and an unusually negative effect of religious extremist perpetrators on casualties.

Below Table 3 we show the mean deaths produced by each cluster. (The total number of deaths in our sample of 181 events was 966; the overall mean is 5.3.) We also show the percentage of events in each cluster that resulted in any deaths. (The mean for the whole sample is 23.8%.)

Cluster A, whose overall profile of net effects is very similar to that of the sample as a whole (compare column 1 to the final column of Table 3), accounts for half the total deaths (505 out of 966 deaths in our sample). However, take a look at Cluster C. This cluster consists of 17 events that had a mean number of deaths of 7.59 killed per event (much higher than the Cluster A mean of 5.26), and 76.5% of the Cluster C events resulted in deaths (compared to 17.7% for Cluster A). By these measures, the Cluster C events are very lethal. The bad news, however, is that the profile of Cluster C not only looks extremely different from that portrayed by the usual regression coefficients (compare column 3 to the final column in Table 3), but the net effects for Cluster C are all tightly close to 0. In other words, unlike Cluster A, the events in Cluster C form a “clump” that was entirely poorly predicted by the usual regression model. The good news, though, is that we can identify such “clumps” by use of our procedures. Looking at the composition of Cluster C events, they tend especially to be events perpetrated by religious extremists who used explosives.

6. In conclusion

We have illustrated a new approach to "turning regression analysis inside out," and we have indicated applications of that approach to the study of CBRN events such as those in the START Center's POICN database.

We have illustrated a "dual" to regression analysis, one in which the key results of the regression (including the coefficients and the predicted values on Y) may be seen to depend on the *cases*. This allows us to use the variables to learn about the cases, whereas the standard use of regression analysis is to make the cases invisible and to discuss only the relations among the variables (see also Shalev 2007 for a critique of the use of regression analysis in the comparative study of welfare states). Unlike many critics of "general linear reality" (Abbot 1988, and, subsequently, many other critics), we do not want to overturn regression modeling. To the contrary, we want to get more out of it. We do so by recognizing foundations of regression modeling in "spaces" similar to those used by practitioners of case-oriented techniques (Ragin, 2008), factor analysis (Shalev, 2007), and correspondence analysis (Le Roux and Rounet, 2004).

With respect to the prediction of CBRN events and their key features, our analysis demonstrates the possibility that standard linear models do well in portraying major regions of the data (such as our Cluster A), while at the same time more than one story about how predictors affect the outcome is necessary in order to describe adequately the data as a whole. We seem poised to be able to identify multiple causal "recipes," and also to identify subsets of cases for which a particular linear model does not perform well. These abilities suggest the value of continuing to work on turning regression modeling inside out.

References

Abbott, Andrew. 1988. "Transcending General Linear Reality." *Sociological Theory* 6 (2):169-186.

Ackerman, Gary A., and Lauren L. Pinson. 2012. "Effectively Managing and Analyzing Empirical Data on CBRN Terrorist Events." Paper Presented at the International Studies Association Annual Convention, San Diego, CA (April).

Asal, Victor, et al. 2012. "Exports of another Type: The Determinants of Interstate Transmission of Terrorism." Paper Presented at the International Studies Association Annual Convention, San Diego (April).

Belsley, David A., Edwin Kuh, and Roy E. Welsch. [1980] 2004. *Regression Diagnostics : Identifying Influential Data and Sources of Collinearity* . Hoboken, N.J.: Wiley.

Breiger, R.L., G.A. Ackerman, V. Asal, D. Melamed, H.B. Milward, R.K. Rethemeyer, and E. Schoon. 2011."Application of a Profile Similarity Methodology for Identifying Terrorist Groups that use Or Pursue CBRN Weapons." Pp. 26-33 in *Social Computing, Behavioral-Cultural Modeling and Prediction*, edited by J. Salerno, S.J. Yang, D. Nau, and S. Chai. Berlin;Heidelberg: Springer.

Gelman, Andrew, and Jennifer Hill. 2007. *Data Analysis using Regression and multilevel/hierarchical Models* . Cambridge ; New York: Cambridge University Press.

Gunst, Richard F., and Robert Lee Mason. 1980. *Regression Analysis and its Application : A Data-Oriented Approach* . New York: M. Dekker.

Hayden, N.K. 2009."Terrifying Landscapes: Understanding Motivations of Non-State Actors to Acquire and/or use Weapons of Mass Destructio**n**." Pp. 163-194 in *Unconventional*

Weapons and International Terrorism: Challenges and New Approaches, edited by M. Ranstorp, and M. Normark. London and New York: Routledge.

Krebs, Valdis. 2001. "Mapping Networks of Terrorist Cells." *Connections* 24 (3):43-52.

Le Roux, Brigitte, and Henry Rouanet. 2004. *Geometric Data Analysis: From Correspondence Analysis to Structured Data Analysis* . Dordrecht: Kluwer.

Melamed, David, Ronald L. Breiger, and Eric Schoon. 2012. "The Duality of Clusters and Statistical Interactions." "Conditional Accept" from *Sociological Methods and Research*

Melamed, D., E. Schoon, R. Breiger, V. Asal, and R.K. Rethemeyer. 2012."Using Organizational Similarity to Identify Statistical Interactions for Improving Situational Awareness of CBRN Activities." Pp. 61-68 in *Social Computing, Behavioral-Cultural Modeling, and Prediction (Lecture Notes in Computer Science 7227)*, edited by S.J. Yang, A.M. Greenberg, and M. Endsley. Berlin; Heidelberg: Springer-Verlag.

Pedahzur, Ami, and Arie Perliger. 2009. *Jewish Terrorism in Israel* . New York: Columbia University Press.

Perliger, Arie, and Ami Pedahzur. 2011. "Social Network Analysis in the Study of Terrorism and Political Violence." *PS: Political Science & Politics* 44 (01):45-50.

Ragin, Charles C. 2008. *Redesigning Social Inquiry: Fuzzy Sets and Beyond* . Chicago and London: University of Chicago Press.

Roberts, Nancy, and Sean F. Everton. 2011. "Strategies for Combating Dark Networks." *Journal of Social Structure* 12 (2):

Rodriguez, Jose A. 2005. "The March 11th Terrorist Network: In its Weakness Lies its Strength."

Schelling, Thomas C. 1960. *The Strategy of Conflict* . Cambridge: Harvard University Press.

Shalev, M. 2007."Limits and Alternatives to Multiple Regression in Comparative Research." Pp. 261-308 in *Comparative Social Research (Symposium on Methodology in Comparative Research),* edited by L. Mjøset, and T.H. Clausen. Elsevier.

Sparrow, Malcolm K. 1991. "The Application of Network Analysis to Criminal Intelligence: An Assessment of the Prospects." *Social Networks* 13 251-274.

START Center (National Consortium for the Study of Terrorism and Responses to Terrorism, University of Maryland). 2012. "POICN Database and Codebook, TANC-D Variables and Traits, V. 8.0."

Tsvetovat, Maksim, and Kathleen M. Carley. 2005. "Structural Knowledge and Success of Anti-Terrorist Activity: The Downside of Structural Equivalence." *Journal of Social Structure* 6 (2):

Table 1: Descriptions of Outcome Variable and of Predictor Variables*

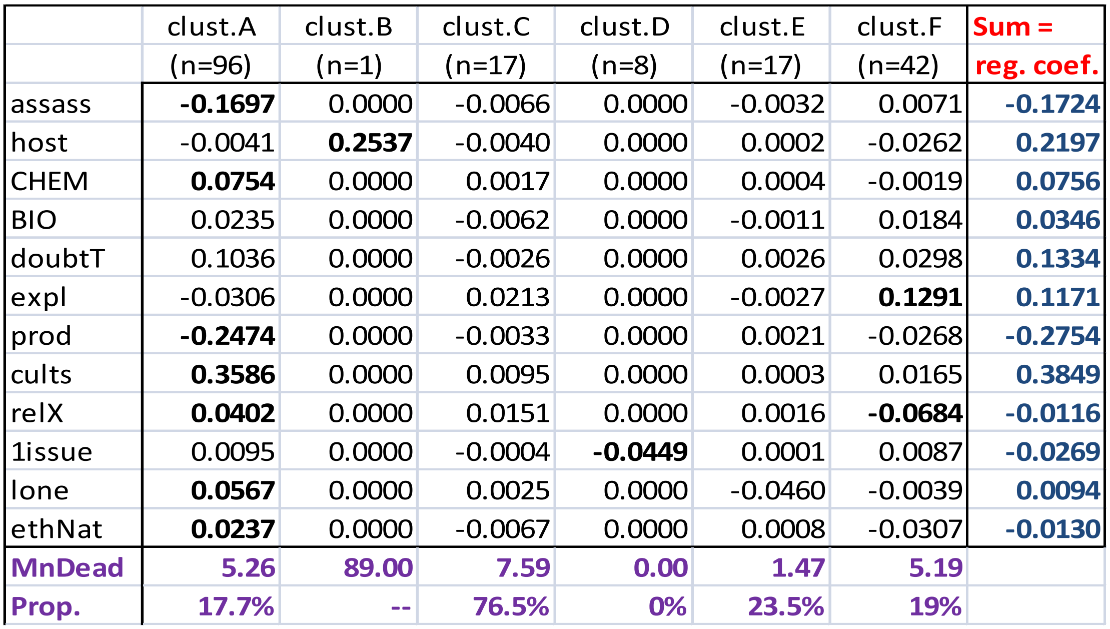
Variable	Brief definition	Mean or Proportion (SE)
Total killed	The number of all individuals who died as a direct result of the CBRN event	5.34 (24.51)
assassination	The event reportedly intended to involve an assassination, an attempted assassination as primary objective	0.138
Hostages	An event whose primary objective is to obtain political or other concessions in return for the release of prisoners (hostages)	0.006
CHEM	The event reportedly involved the use of a toxic chemical agent	0.884
BIO	The event reportedly involved the use of a biological agent	0.088
doubtT	Reservation, in the eyes of POICN-D analysts, that the event in question is truly terrorism	0.431
explosive	The CBRN attack used an explosive.	0.171
production	Based on source information, the event involves the perpetrator producing some level of the agent "in house"	0.166
cult	Perpetrators included any type of cult, including religious groups. Distinguished from other organizations by their authoritarian internal social control mechanisms rather than by their specific theologies or ideologies	0.144
relX	All groups operating in the name of religion that do not fall under “religious cult”	0.094
1issue	A perpetrator addresses a single issue, such as environmentalism	0.044
loner	Individuals who are not operationally linked to any larger groups	0.094
ethNat	Perpetrator place greater importance on descent/heredity than political borders; may be pursuing sovereignty or additional rights	0.232

*For complete description, see the codebook (START Center, 2008)

Table 2. Regression of Total Killed on predictor variables

Xi	Est	SE	t	Pr(>|t|)
assass	-0.172	0.09	-1.97	*
host	0.220	0.07	3.02	**
CHEM	0.076	0.11	0.72	
BIO	0.035	0.11	0.31	
doubtT	0.133	0.08	1.71	.
expl	0.117	0.09	1.26	
prod	-0.275	0.12	-2.39	*
cults	0.385	0.12	3.15	**
relX	-0.012	0.09	-0.13	
1issue	-0.027	0.07	-0.36	
lone	0.009	0.08	0.12	
ethNat	-0.013	0.08	-0.16	

Table 3. The usual regression coefficients (see Table 2) are sums across clusters of cases*



	clust.A	clust.B	clust.C	clust.D	clust.E	clust.F	Sum =
	(n=96)	(n=1)	(n=17)	(n=8)	(n=17)	(n=42)	reg. coef.
assass	**-0.1697**	0.0000	-0.0066	0.0000	-0.0032	0.0071	**-0.1724**
host	-0.0041	**0.2537**	-0.0040	0.0000	0.0002	-0.0262	**0.2197**
CHEM	**0.0754**	0.0000	0.0017	0.0000	0.0004	-0.0019	**0.0756**
BIO	0.0235	0.0000	-0.0062	0.0000	-0.0011	0.0184	**0.0346**
doubtT	0.1036	0.0000	-0.0026	0.0000	0.0026	0.0298	**0.1334**
expl	-0.0306	0.0000	0.0213	0.0000	-0.0027	**0.1291**	**0.1171**
prod	**-0.2474**	0.0000	-0.0033	0.0000	0.0021	-0.0268	**-0.2754**
cults	**0.3586**	0.0000	0.0095	0.0000	0.0003	0.0165	**0.3849**
relX	**0.0402**	0.0000	0.0151	0.0000	0.0016	**-0.0684**	**-0.0116**
1issue	0.0095	0.0000	-0.0004	**-0.0449**	0.0001	0.0087	**-0.0269**
lone	**0.0567**	0.0000	0.0025	0.0000	-0.0460	-0.0039	**0.0094**
ethNat	**0.0237**	0.0000	-0.0067	0.0000	0.0008	-0.0307	**-0.0130**
MnDead	**5.26**	**89.00**	**7.59**	**0.00**	**1.47**	**5.19**	
Prop.	**17.7%**	**--**	**76.5%**	**0%**	**23.5%**	**19%**	

*Notice that the sum across any row (i.e., the sum across the clusters of cases for a particular variable) yields the regression coefficient (see Table 2) for that variable.

In parentheses under cluster name (e.g., “n=96”) are the number of events comprising each cluster.

Below the table, “MnDead” is the mean number of deaths within each cluster (respectively), and “Prop.” Is the proportion of events in each cluster with at least one death.

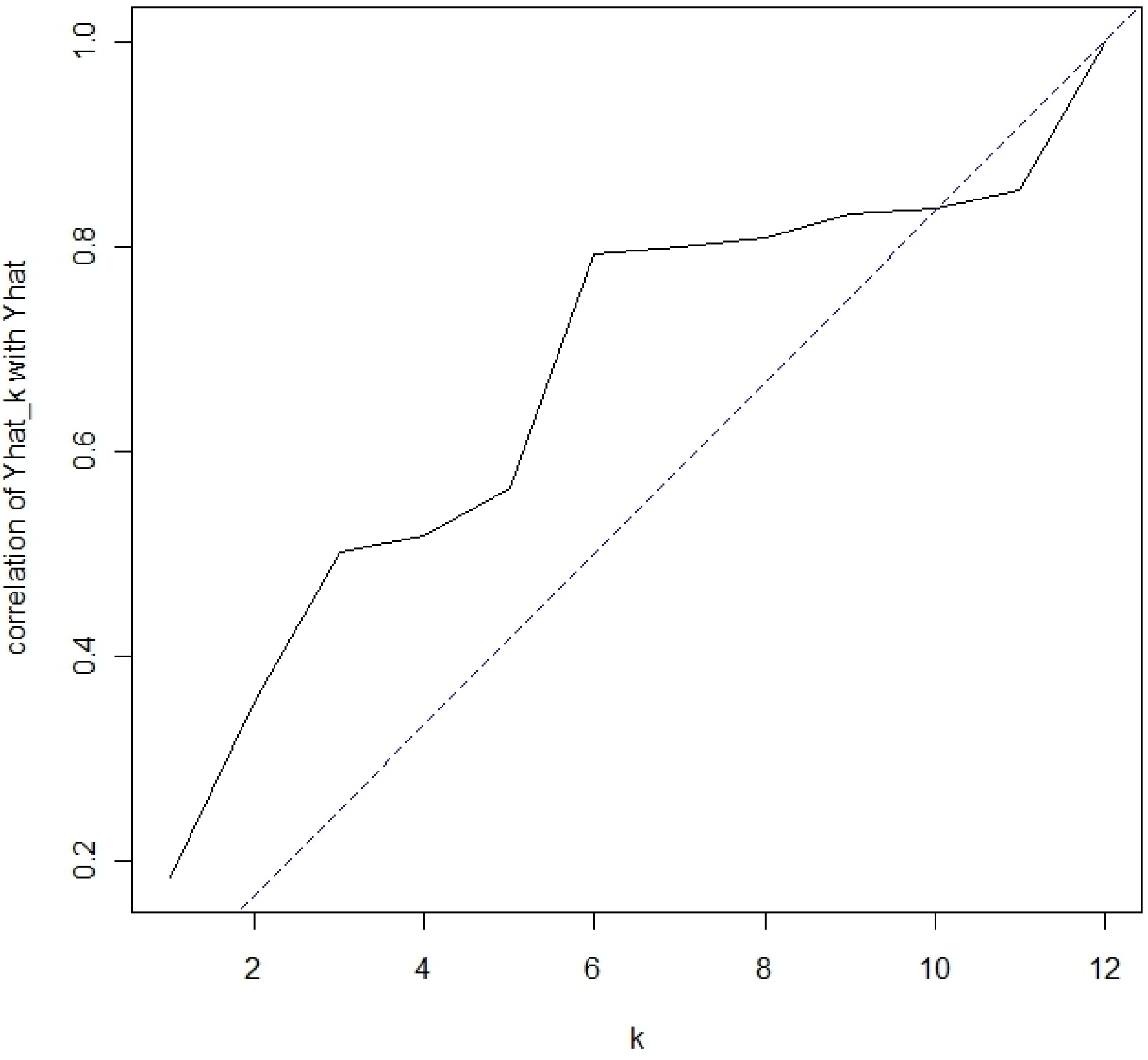


Figure 1. Correlation of $\hat{Y}_{\mathrm{k}}$ with $\hat{Y}$, for k = 1, …, 12

$$\hat{Y} = \mathbf{U}[,1\!:\!k]\ t(\mathbf{U}[,1\!:\!k])\ \mathbf{y}$$

where **U** comes from the SVD of data matrix **X**, and **y** is the observed outcome variable.

The dashed line shows the cumulation of (1/12) shares of the total.

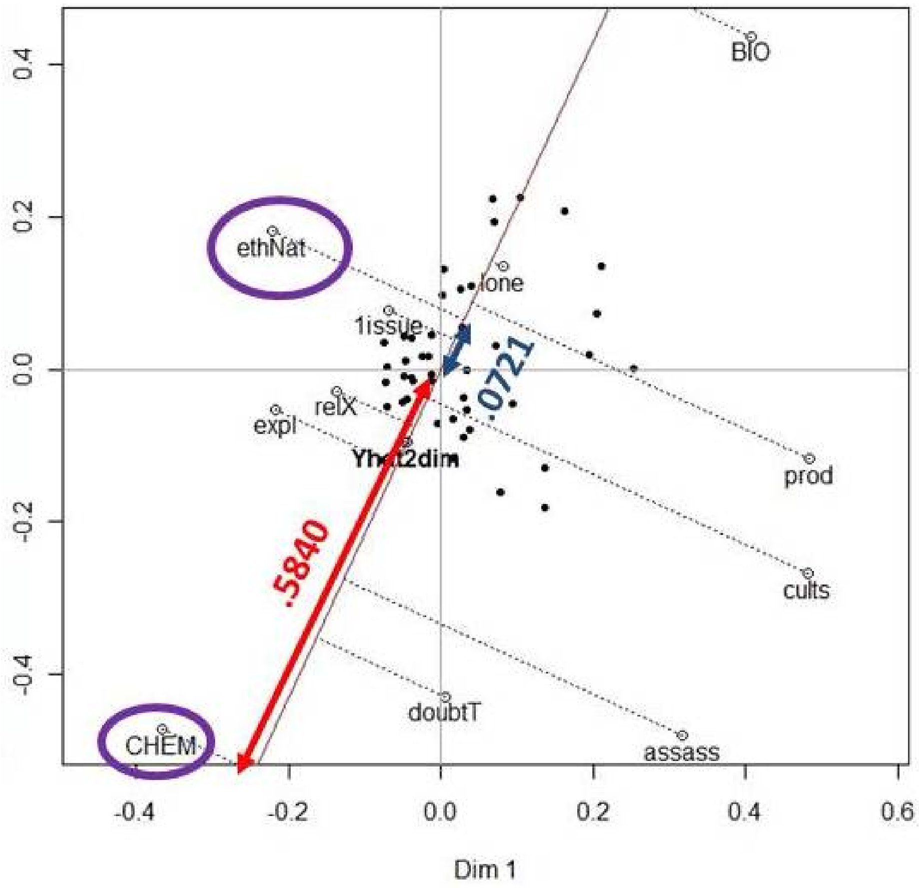


Figure 2: A graph of the first two dimensions of **U** (cases) and **V** (points). Red line (marked .5840) is distance of projection of CHEM from the origin. Blue line (marked .0721) is distance of projection of "ethNat" from the origin. Ratios of these distances (.5840 / .0721 = 8.0971) are identical to ratios of the corresponding regression coefficients (in two dimensions: .0626 / -.00773 = -8.0971). The same holds true for all pairs of regression coefficients.

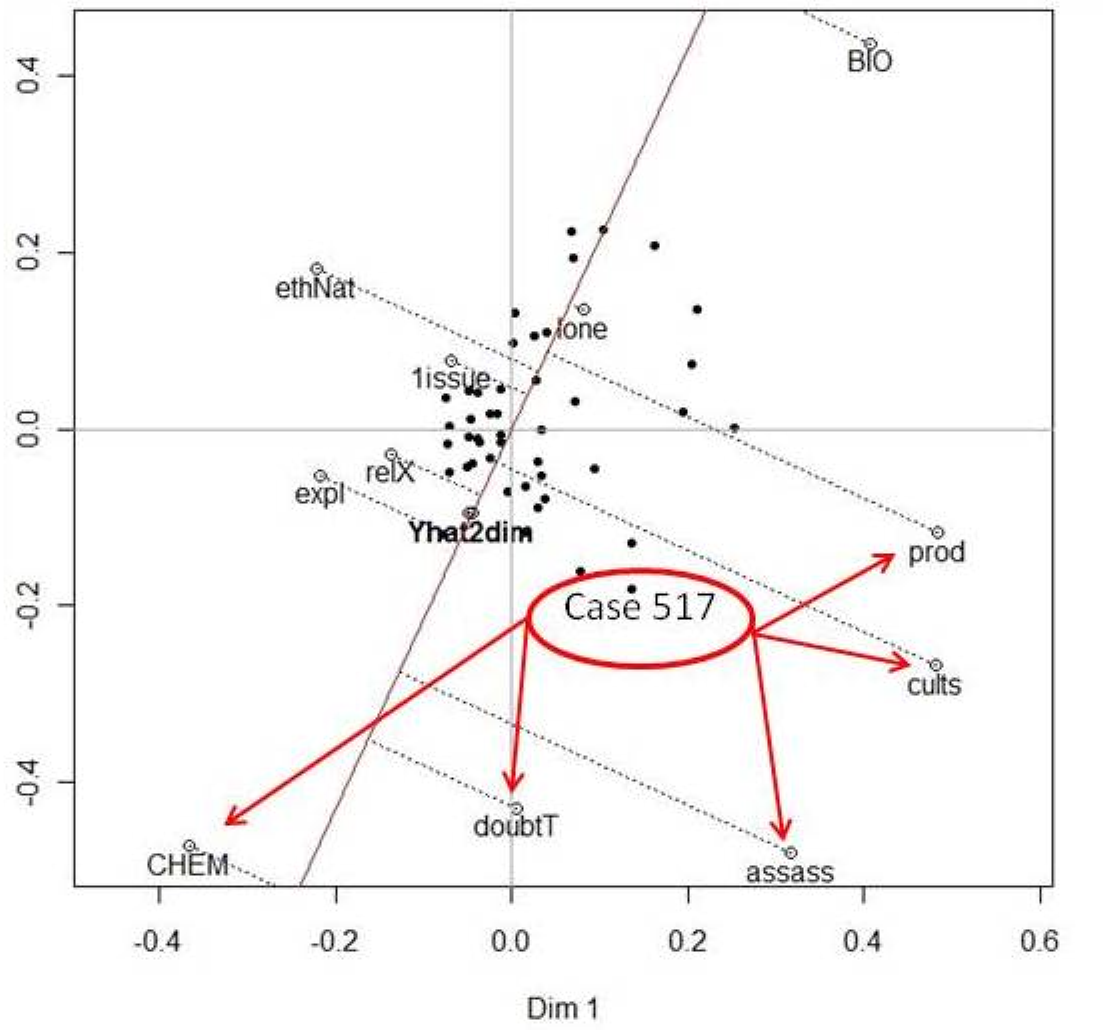


Figure 3. The position of Case 517. The narrative of this case is (in part): ": "In late November [1994], Aum Shinrikyo cult members … attempted to assassinate Noboru Mizuno … [who] was helping deserters from Aum … escape…. [One of four attackers] tried to spray VX nerve agent, from a syringe, on Mizuno's neck…." **Properties** of case 517 include: **CHEM**, **doubtT**, **assass**, **cults**, **prod** [the agent was produced "in-house"].

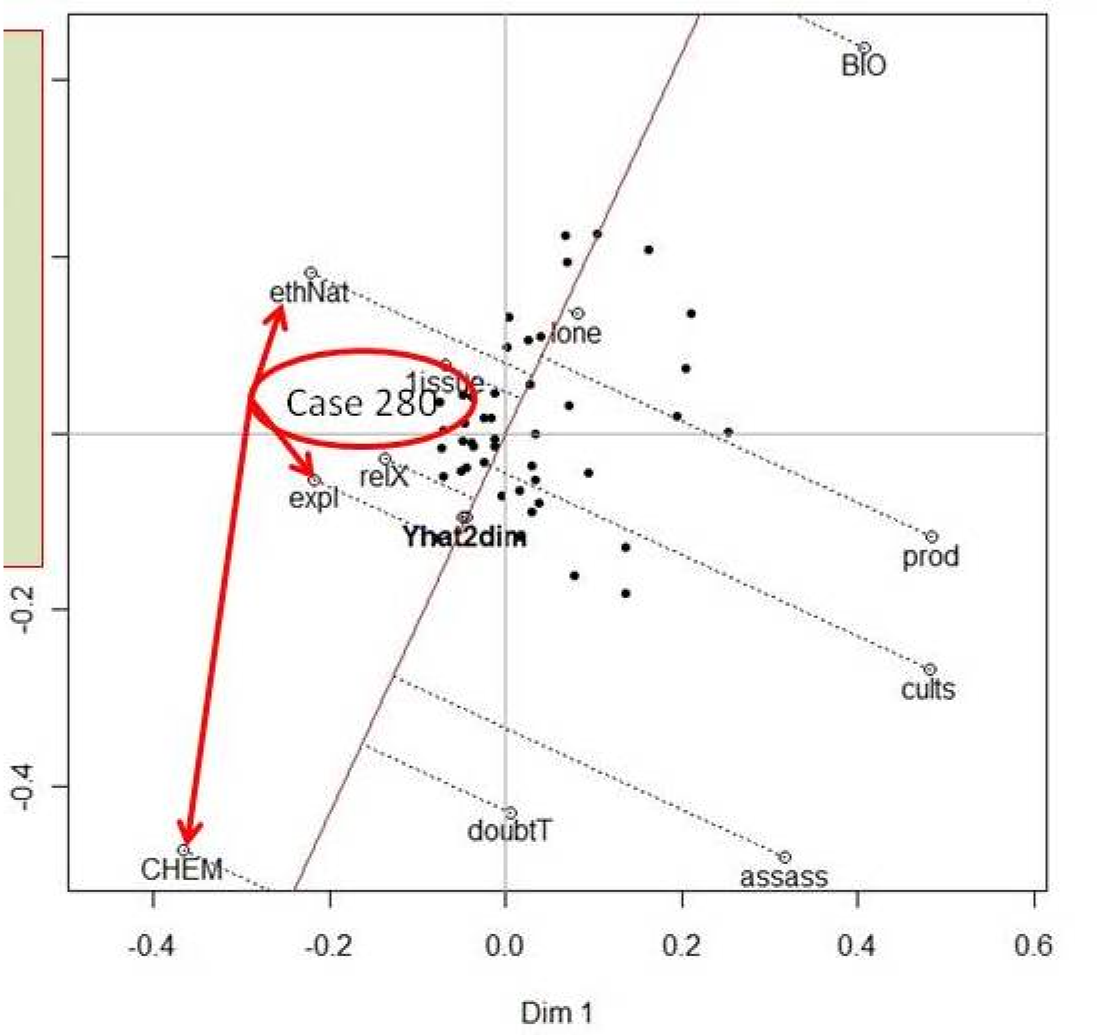


Figure 4. The position of Case 280. The narrative of this case is (in part): “On 15 July 2003, a homemade explosive device was found placed on tanks of acetone and phenol at … [a chemical factory] in Saratov, Russia…. Chechen Rebels were suspected to have placed the device … to intimidate Moscow….” **Properties** of case 280 include: **CHEM**, **expl, ethNat**.

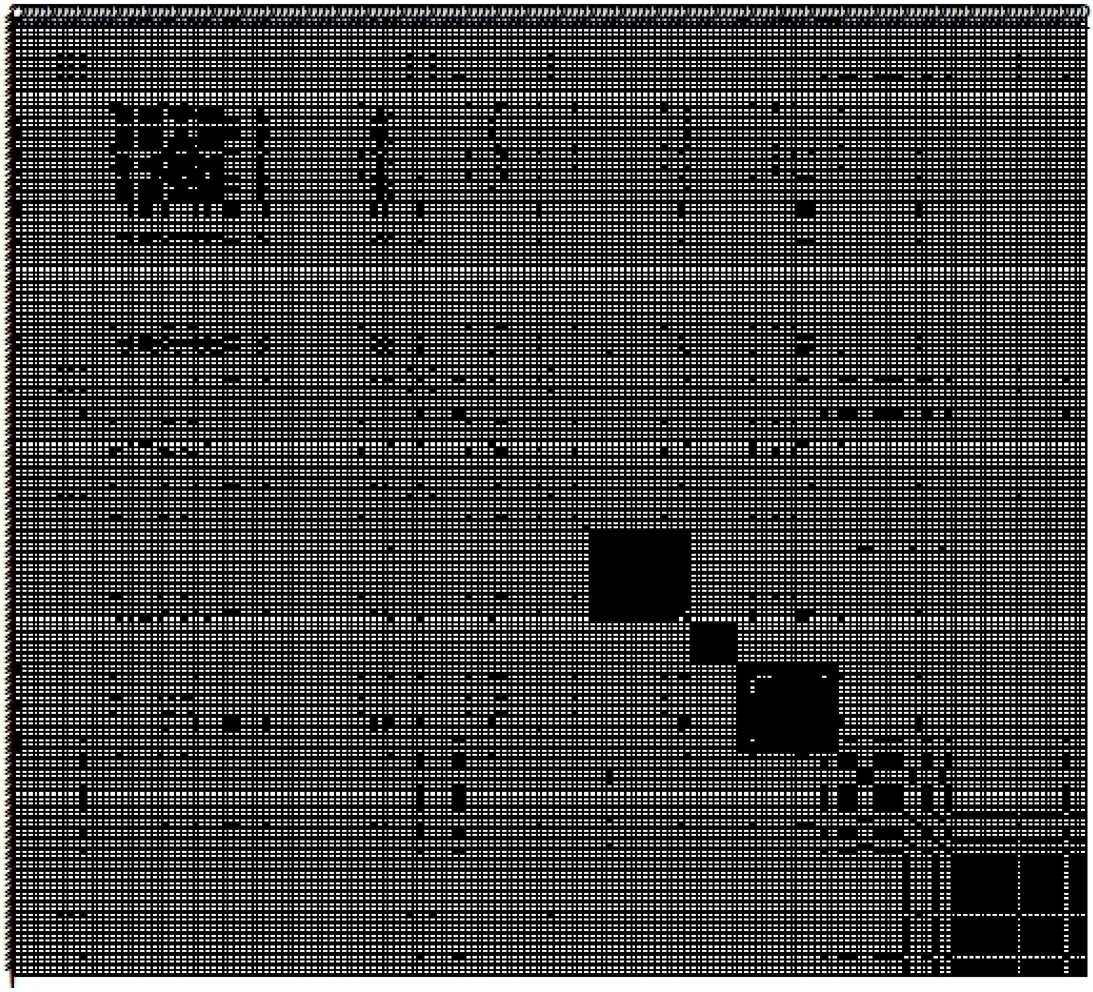


Figure 5. Clustering of the network among the cases (UU^T; eq. 7). We use full numeric information in performing the cluster analysis. Purely to aid visualization, above shows the clustering applied to values greater than an arbitrary cutoff value.