Igor Mandel
Telmar Group Inc., 711 Third av., New York, NY, imandel@telmar.com

Fusion and causal analysis in big marketing data sets

Abstract

Since so many marketing studies reflect differing aspects of consumer behavior, there is today a critical need for data fusion. But fusion can be challenging, especially when we are talking about fusing thousands of variables. In this article, we present an approach which solves this problem in a highly efficient manner. Another problem is in applying causal types of models (as opposed to traditional statistical ones) for the analysis of complex marketing data. A new, intrinsic probabilities, approach is proposed and compared with others.

1. Fusion in big data sets

The purpose of ascription (fusion) is to merge information of two datasets into one, in a such a way, that external criteria are satisfied.

Let assume we have two data sets: A, the source of information to be ascribed (source) and B (hub), with different numbers of respondents.. They both have the same subset of common binary (usually demographic) variables X. Data set A has also Y variables, binary and/or numerical (usually media or brand related). The goal is to create dataset C, which contains the entire dataset B together with, ascribed to each respondent, values of variables Y. Ascription should satisfy two conditions:

1) Compositions (and/or typical numerical values) of Y variables, for each value of X variables in C, should be close to ones in B

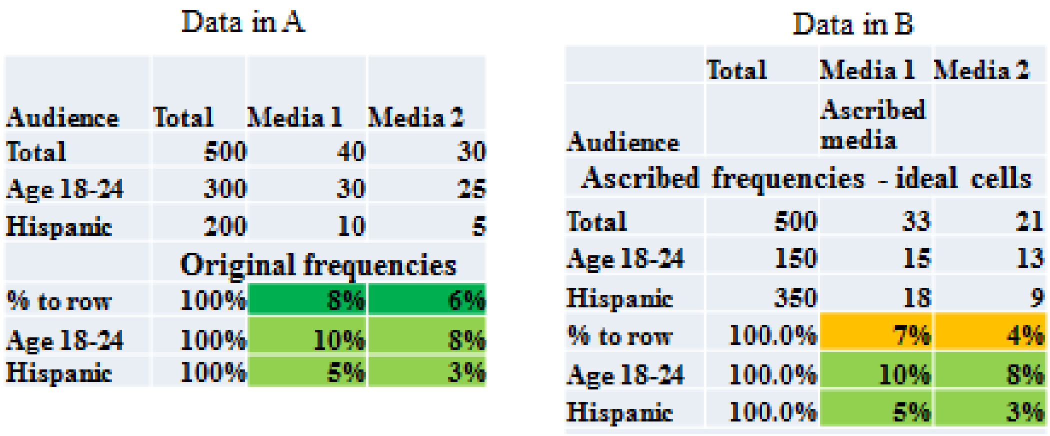
2) Correlations between Y variables in C should be close to correlations in B.

These conditions, especially 2), in the case of many Y variables, are very hard to satisfy for understandable reasons. It's enough to note that the number of correlations between variables is proportional to the square of number of variables; a typical data set with about 1,000 to 5,000 media vehicles gives several millions of pairs. But correlation in media planning is not just of academic interest; for media vehicles, it translates as duplication (when the same person is exposed to advertising in several media channels), that can directly affect media budget allocations, etc.

For that last reason alone, the traditional methods of making some kind of models for Y variables on data A and then plugging them in B, using the common variables X (as, for example, in [1], do not appear to be sufficient. Even if problem 1) could be solved well, it doesn't help to solve problem 2). But usually even requirement 1) cannot be satisfied with a direct modeling approach. The reason is that relations between all Y and X variables on A should be strong enough to make a good prediction – which is very rare, at least in my experience.

There is another problem with practical fusion. The farther structures of X variables in two datasets are from each other – the worse the fusion results are, no matter how good the prediction from X to Y is in a source. This is easy to demonstrate with a numerical example, as shown in Figure 1.

On the left pan, data set A, there is a fragment of source data: it has original observed values for two demographic variables and two media variables. In data set B, demographic variables are observed, but media variables – ascribed. The demographic structure in B is different from one in A: Hispanic represents 70% in B vs. 40% in A. The purpose of ascription was to make all proportions of data B, as close as possible to ones in A – but in what sense?



Data in A

Audience	Total	Media 1	Media 2
Total	500	40	30
Age 18-24	300	30	25
Hispanic	200	10	5
	Original frequencies		
% to row	100%	8%	6%
Age 18-24	100%	10%	8%
Hispanic	100%	5%	3%

Data in B

	Total	Media 1	Media 2
Audience		Ascribed media	
Ascribed frequencies - ideal cells			
Total	500	33	21
Age 18-24	150	15	13
Hispanic	350	18	9
% to row	100.0%	7%	4%
Age 18-24	100.0%	10%	8%
Hispanic	100.0%	5%	3%

Fig. 1. Effect of demographical structure differences to results of fusion

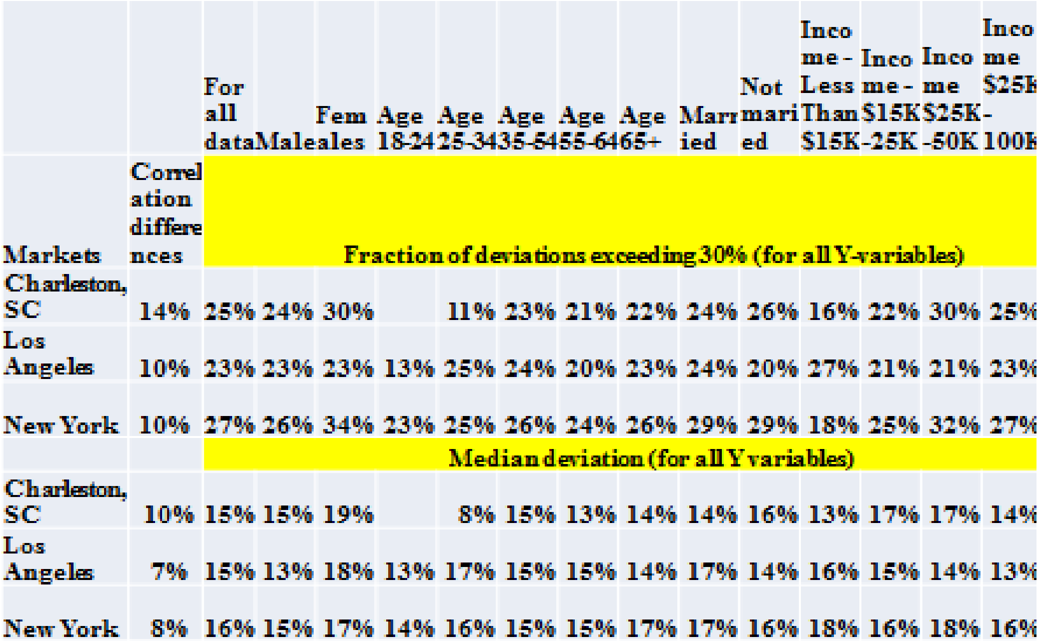
If one tries to minimize the particular cells difference for percentages to rows, which are in light green in A, and have the ideal fusion algorithm, which makes frequencies of fused data the same in B –then inevitably totals will be different: (brown in B vs. dark green in A). Similarly, if one minimizes the totals differences and fit it perfectly – the cells will be different.

So, when demo structures in two data sets are different – it is impossible to make a very good fusion, even using ideal algorithms. And this is a very typical situation for many marketing data sets: their demographic bases are different. Of course, one can make a compromise between totals and particular cells, but then it is not clear which of those are more important, keeping in mind the final goal of a fusion, and how to coordinate these two aspects.

Yet another problem is that dealing with many thousands of Y variables, and controlling for tens of X variables, create so many combinations (especially if taking into account correlations), that computational problems may just be prohibitive.

In order to address these and other issues, a new approach was proposed. In essence it is a specific way of **cloning** respondents from A into B in a most plausible fashion. It is different from predictive modeling in one important aspect: all Y variables are considered “in bulk”, not modeled separately; the idea is to imitate Y behavior in general, rather than understand why it happens for each variable (what modeling usually puts as a goal). The details of algorithms are proprietary for Telmar Inc., but the results are rather impressive.

One hundred thousand respondents in data set A were fused with 3.1 m respondents in Centab, a Telmar demographic data base [2]; the two data sets had significantly different demographic compositions, which as shown below, worsens results, but the level of errors is still good, keeping in mind that around 1,000 variables were ascribed (Fig.2).



		For all data	Male	Fem eales	Age 18-24	Age 25-34	Age 35-54	Age 55-64	Age 65+	Marr ied	Not mari ed	Inco me - Less Than $15K	Inco me - $15K -25K	Inco me $25K -50K	Inco me $25K - 100K
Markets	Correl ation differe nces	Fraction of deviations exceeding 30% (for all Y-variables)													
Charleston, SC	14%	25%	24%	30%		11%	23%	21%	22%	24%	26%	16%	22%	30%	25%
Los Angeles	10%	23%	23%	23%	13%	25%	24%	20%	23%	24%	20%	27%	21%	21%	23%
New York	10%	27%	26%	34%	23%	25%	26%	24%	26%	29%	29%	18%	25%	32%	27%
		Median deviation (for all Y variables)													
Charleston, SC	10%	15%	15%	19%		8%	15%	13%	14%	14%	16%	13%	17%	17%	14%
Los Angeles	7%	15%	13%	18%	13%	17%	15%	15%	14%	17%	14%	16%	15%	14%	13%
New York	8%	16%	15%	17%	14%	16%	15%	15%	17%	17%	16%	18%	16%	18%	16%

Fig. 2. Relative errors of ascription for 1000 Y variables

Each value in a table was calculated on all vehicles – say, value for Charleston/ Females means that median value of deviation in frequencies between ascribed and observed data for Females in Charleston for all 1000 vehicles is 19%. Deviation was calculated as the absolute difference between two frequencies divided by the frequency of the observed data. Of course, with this formula, small frequencies in A create, in general, large deviations, although they may not mean much in a practical sense. For example, if in dataset A, female readers of Vogue represented 0.5% of all females, while 1.5% in ascribed dataset B, then the statistic will be 300%. But for media planning both figures are considered just "small". But even with this caveat, deviations are not that large.

Errors for correlations (the first column) were calculated based on frequencies of interception between the vehicles. It is of special importance that are also small. All errors are about the same for large (like New York) and small (like Charleston) cities, what is also remarkable. Computational time for fusion was about 2 hours.

2. Causal analysis – the intrinsic probabilities approach

There is no place here to discuss numerous problems, controversies, and drawbacks in these approaches; it was discussed in length in [4,5] and other works.

The past two decades have witnessed a burst of works about the causality problem, and causality has been considered in statistical literature from different angles. The main approaches under development were simultaneous structural equations (historically the first, derived from works of S. Wright in 20-s); potential outcomes, proposed methodologically by E. Neiman in 20-s and technically developed by D. Rubin and others from 70-s; concept of do-operators and associated with them acyclic graphs, developed by J. Pearl and colleagues started from 80-s and tightly related with Bayesian networks theory. Some methods and approaches were developed by J. Robins, W. Dawid, I.Mansky and other researchers; a little bit aside stays "Granger's causality" theory. There is an opinion, defended prominently by J. Pearl [3], that almost all of these approaches in fact talk about the same things, but use different terms and stress different aspects of the problem. I'm not sure it is rightly justified or not, but if it is – what follows would thus get even more indirect support from this observation..

I cannot here consider all the literature even at glance, but there is a strange impression from all that I have looked at: practically **no authors introduce a clear definition** of what the cause is and what is, respectively, the topic of all these studies. Maybe, the vagueness of the term "causality" is the main reason why some prominent scientists do not really consider it seriously or use it reluctantly. One of these is L. Zadeh, inventor of fuzzy logic, who noticed once that he does not see a way to define causality strongly, and for that reason, cannot see a possible theory [4]. In one of the badly determined examples he brings the following: *a friend of mine calls me up on the telephone and asks me to drive over and visit him. While driving over, I ignore a stop sign and drive through an intersection. Another driver hits me, and I die. Who caused my death?*

Out of all four chained events which "caused the death" – friend's call, driving a car, ignoring a red light and being hit by the car - the argument goes, one cannot formally separate "real causes" from "fictitious" ones, and so a theory is not possible (L. Zadeh brought other arguments, too). As J. Pearle remarked on that, it leaves him (Pearl) "scientifically unmoved" http://bayes.cs.ucla.edu/BOOK-2K/hautaniemi.html – i.e. he will continue working on the theory, but by a weird coincidence, he still doesn't say what exactly to do in situations like that. His "do" operator fixes values of certain variables in order to see how others behave, but this very procedure, as will be shown, does not address the Zadeh's issues.

2.1 Individual and statistical causes.

Causes could be looked at under different angles, and mixing these is, maybe, the main source of confusion in literature. Let's consider here briefly only two of these aspects: relationship between causes of **individual event** and causes in a **statistical sense** (where there are many individuals or events). There are other separation lines, like causes of very particular **events** (breaks in engineering) vs. causes of **processes**, as in social life, etc., but I cannot consider them all here.

If one takes any individual situation where the question "why?" (ie. appeal for causal explanation) rises, it will be always like in Zadeh's example given above. It could be formally presented as chain of events causing one another, where on each event point there is a chance for potential "No", i.e. for assumption that this given event didn’t happen and all further part of a chain would not be materialized. In the provided example:

a. if friend didn’t call – he will not drive;
b. if he called, but he was not in hurry on a way – he will not be hit by a car;
c. if car’s driver was more experienced – he will not be hit by the car;
d. if car was driving with less speed – the crash will be not fatal, etc.

Technically, it is a graph with many nodes going to eternity at the bottom and to death at the top, where from each node there are two or more imaginable options, out of which just one was real and all other “would be” but were not – see Fig.3. At each important point there is an opportunity to change something (do not react to call, do not drive, etc.), but they all are ignored in a real stream of life. What is important – each lost opportunity immediately generates a chance for another, this time illusory application of certain force, i.e. making action, which in turn, would be mentally prolonged even further (with cutting other illusory opportunities) – and, if desired, the whole chain could easily end up in the same terminal point B. In our case – a driver could not ignore a stop sign, but still be dead for many other possible reasons.

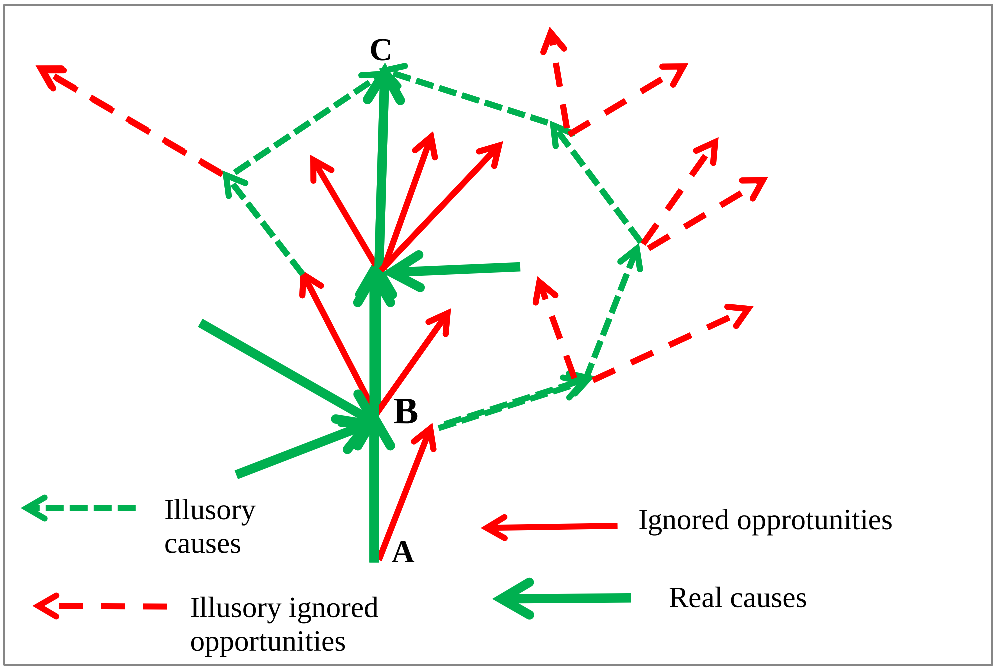


Fig. 3. Graph of real and illusory causes and opportunities for event C

The point B on a chart shows, that two other real causes (besides of one originated from A) affected decision to be made at this moment (A is a “friend’s call”, B could be “decision to drive”, and the two side causes could be “I have a car” and “I have time to go”). The traditional causal theory presumes, that causes like that not only work, but work on statistical level as well, i.e. for many situations (see below). However, it is quite clear, that there are immeasurable number of similar “real causes”, and what is most important, they are applied to each point of the chain. It is absolutely impossible to discern them from many observations and from many trees like that. They in fact would be an “***illusory causes of the second degree***” – in that case not because they have been rejected as red arrows, but because they are deeply enrooted in any decision and for that reason undistinguishable from it. In other words, the two reds from B were made after considerations (subconscious, for this matter) of all these thick greens – and for that reasons these greens are irrelevant, even on individual level. Much more so on a level of many trees.

This picture is in fact nothing more than an ***evolutionary tree***, applied to each event caused by “all history”. If understood as such, all possible theories of evolutionary development could be applied into causal analysis (see the excellent survey of economic and social life from this angle in [4]). The only problem with evolution is, however, that it is, strictly speaking, not to be predicted and even modeled at all. The only plausible way to deal with it is to use a certain type of agent based modeling imitation, which, in turn, has many internal problems [6].

On a side note, trees like those on Fig.3 are the basis for different tales or “alternative histories” – with omission of the fact, that if you assume just one red arrow as “materialized opportunity” - like “would the first World War happen, had Gavrilo Princip not killed Archduke Franz Ferdinand?”- you cannot really imagine following broken greens and even less broken red arrows. Usually, the imagination of authors goes just to the first fork.

So, coming back to formal modeling of causality on an individual level, the first problem is to find a **stopping criterion** – where to break a chain in order to make it meaningful. It was exactly the problem which Zadeh thought to be formally unsolvable, and, strongly speaking, this argument is impossible to deny, unless one makes a forced decision – say, to cut always after the first level. The most interesting illustrations to that difficulty give many courts decisions, which, supposedly, should reveal “the real causes” and charge those who are guilty (in fact, causality should be the only business of the entire justice system). When, for instance, a lady bought hot coffee in McDonalds, put it between her knees, started to drive and then after spilling it famously won a case against McDonalds (“*cause of the burnt skin is that coffee was hot when she bought it*”) – this type of causal relations would be really hard to set as plausible in any not insane framework, but it is yet a “legal cause” to pay her “compensation”. In this case all parts of the chain are for certain reasons ignored, and the stopping criterion was “*McDonald’s fault in making coffee hot*”. But it could easily go further (“why people drink hot coffee”), or shorter (“why she put it between her knees”), what will create a different legal output, from “drinking hot coffee is illegal” to “plaintiff’ claim is denied as not grounded” etc.

Now let’s imagine that we have many “individual stories’ like that, i.e. we try to understand a typical statistical situation when some kind of repetitive, not unique events, take place. The first question to be asked is: what is on the top of these charts (with any depth)? It is not trivial. For the police department it is clear that we should collect only cases with death at the end (continuing Zadeh’ example) – otherwise we lose the point of the study. If we follow this logic, we’ll have, say, thousands of charts with the same result – death in a car accident – and very different patterns of the histories behind them. It is quite clear that the event “friend’s call” will appear in these histories very rarely if ever, but, say “running on a street” – much more often. But can such data potentially give us? Would it provide us with the causal explanations, even if the trees are deep enough and we analyzed them carefully?

Many people would say “yes”. This is, for example, how medical statistics works: to the question what are causes of death, it lists all “prima causes”, like “*cancer”, “stroke”* and so on and the distribution of these causes gives us a picture of the phenomena. Going to the next level, they may add (although with bigger difficulties) that “cancer” could be “lung”, ‘kidney”, etc., but logic does not change with this – it’s still ***listing the causes***, ***but not revealing the mechanism*** of the event. As we reached the point “why lung cancer happened” – then immediately this logic fails and a different story began. The same is typical to almost all technical problems: no one cares about normal process (no one asks question “why car drives?”), but everyone bothers when something happens and needs fixing. Again, if, say, “*lamp doesn’t work*”, the list of possible causes could include “a broken bulb”, “an unplugged cable” and so on but at certain point on a causes tree “explanations” like that became meaningless (just try to systemize “why cable was unplugged?”).

At this very point we reach a situation when collection of data only about events of interest becomes not sufficient. In order to understand why “cables was unplugged” or “lung cancer happened” we need to compare these cases with others, where these target events didn’t take place. Technically, it means, that instead of having Y variable (the outcome) with one value only (say, 1), we have to go to the situation when Y could be either 1 or 0. But what does it really mean if we are still within our set of individual models? For police department, it means that not only “***death cases***” are considered, but “***non-deaths***”, too. Which ones? What kind of history should we collect about these “cases of being alive”? The problem immediately becomes uncertain.

It is clear, that the deeper we go into the chain, the less meaningful our data becomes. If one collects data about “all those who got calls from friend with request to meet her" – it would be a nightmare. If only about those who "run on a street" – it would not be much easier. If about those who were hit by the car, but survived – than it becomes more practical (at least because these cases are usually reported to the police).But the price paid for application of this stopping criterion would be very high: we do not consider millions of these who really were in dangerous situation (run over the street but was not hit). If we don't count for these, we, respectively exclude from consideration such possible measures to avoid accidents as to block the pedestrians from crossing the road. But it was only one of the very many nodes in a set of collected scenarios of deaths. All others are, maybe, not less important – but we really do not know anything about that, for we never are able to collect such a vast data to please our curiosity for "the full causal explanations" on all available trees. Here is one of the very big problems in setting the observational or even experimental study: a ***huge asymmetry*** between "target values" and "non-target outcomes". It is never clear what to put into "control sample" against, say, "Mercedes buyers" – all rich? All urban residents? Whole population? But depending on the answer the whole model will be completely different.

And here we enter the interesting area of the building of scientific paradigms. In real life, people almost always limit themselves just by "prime reasons/causes", like "he died from the cancer", "the cable was unplugged", etc. The whole machinery of the statistical learning also follows the same paradigm: trying to estimate the effect of direct "causes" to the known Y variable. But special causality literature mainly contains models, which consider mutual effects of different variables to each other, i.e. reflecting this eternal chain of events at rather arbitrary cutting points. Ironically, it very often lacks any temporal component – a necessary ingredient of causality in individual modeling, but nevertheless, "circles", "confounders", "counterfactuals" are typically employed slang in a causal literature [3,5,6,7]. But all these terms are just shadows of the individual graphs, transplanted without the reasons into a statistical situation. From the simple consideration of individual causes is clear, that any trees have in fact illusory character – how could it be otherwise for many observations? They may be analyzed via "what-if" techniques (with big difficulties anyway), but for modeling - only way to consider statistical repetitive events is cut off the all leaves and look into repetition of the direct predecessors – materialized causes of the target events.

Why this shift in paradigms happened? Does it mean, that "direct causes analysis" is already a passed stage in science and statistics knows how to deal with it? Not at all – in any textbook one may find a statement that "correlation is not causation", after which authors very smoothly go into models having **only** direct factors (causes) like regression and do not mention causes anymore, unlike in specific causal literature. As a result, it seems there is a strange **gray area**: analysis of direct causes "of the first order" is not really considered with sufficient depth. Where these first order causes are analyzed – they are not called causes, where causes are studied – there is no special treatment for the direct first order. So, I would start with this "simple" situation, when one supposes that some direct causes are known (like independent variables in regression) and needs to make causal, not associational model. And the main conclusion from considering the individual models is that in statistical sense is it impossible and/or impractical to follow all individual trees. The only thing one can do is to consider the probabilities (frequencies) of the target appearance for each variable value. It is a basis of the proposed approach.

2.3 Causal intrinsic probabilities

Let us consider the simplest possible situation when we have a data set containing one variable X – supposedly cause for the variable Y. Let's further assume X and Y are binary, with Y=1 meaning that effect of interest happened. First, let's try to answer the question, which is rarely posed: in what sense we agree that X caused Y? There are several possible situations.

1. X could be a condition, under which Y=1 is always occurring. In this case the cause is called **sufficient** for the effect. The examples are "unplugged cables" for not working lamp and other similar situations, very typical for technical understanding of causality. In this case if X=1, Y always has value 1, but all other combinations are also possible.

2. Y could be such that it doesn't appear without X. The examples could be situations like "water doesn't boil if temperature is below 100 degrees C oat sea level"; "one cannot run for President of US, if she was not born in the US" and so on. Causes like that are called **necessary** for the effect. In this case, if Y=1, X always has value 1, but all other combinations are possible.

3. Up to this point, everyone agreed about these two types of causes. But in reality these two types of causes ither do not exist or attract no interest due to their triviality. The problems arise when one is talking about causes which are **neither sufficient nor necessary**, which represent a lion share of all causes in real life which are important to recognize. Some call them "contributory" in a sense that presence of the causes contributes into appearance of the effect. Some distinguish them from "conditions", which are not causes themselves, but still important to set up the correct framing for causal analysis.

I would not go into details of these debates, but instead propose the following **definition**: ***a contributory cause is a certain combination of circumstances in universe, which produces constant effect within time and space of the studied phenomena***. Let's clarify what all these words mean on simple examples.

a) If any case of “running over the street” means that you are hit by the car – the running would be considered as sufficient cause for being hit. In fact, since but small fraction of cases of running yield this result – one may say, that ratio “hit/running” is a measure of contributory cause intensity.

b) If “percent of mathematicians” among men is, say, 3 times higher than among women, “being a man” is a contributory cause of “being a mathematician”. However, “being a woman” also contribute into the same effect – so, it is also a cause, but with smaller “intensity measure”.

c) If one knows that only people having income more than one million a year could afford to buy a Ferrari, “being in 1 m plus income category” determines a contributory cause for car’s buyers, but having lower income does not.

From these examples follow several features, what contributory causes have.

1. Each **cause works as an “independent entity**”. This is a principle statement. It means, we assume some “intrinsic feature” in each particular “combination of circumstances” such that the effect is always the same. Of course, “always” here is understood within the local content of a particular study and depends on study design – but it is true for any statistical analysis and is not anything unusual. What is unusual though – that “yields” of each cause are somehow internally inherited for the situation which represent a cause. If one thinks that combining several causes permanently creates different yields – these combined causes should be explicitly named and studied separately. If, for instance, one thinks that yield of mathematicians is different for “men” and for “men from wealthy families” – than another variable, “wealthy men”, should be introduced and measured. In that case it is not necessarily, that, say, weighted average yields for wealthy and non-wealthy men would be equal to one for men. The reason is, a new combination “wealthy * men” could create some ***synergetic*** effect, which would increase or decrease yield of any of its components – but we don’t consider it here. In short: one cause – one expected outcome (yield) from it.

2. Each **cause is associated not with a whole variable** (like “gender” in example b) above), but only with **one level** of the variable. One yield is for men, another for women; one is for “running over the street”, another (if any – see below) – for not running, etc. It is also quite fundamental requirement and at the same time it is in full accordance with common use of causal language in everyday life. One cannot say “gender causes mathematical occupation”, but one may say “if you are men, you are more likely to be a mathematician”. This simple and seemingly obvious statement makes however a principal detachment from traditional statistical way of doing analysis: one should look at these “grades related yields” rather than on coefficients of general “association” (or regression), linking “gender” and outcome with one shot. It produces respectively new techniques, as will be shown shortly.

3. **Causes as "physical forces" are not distinguishable from causes as "conditions of life**". Instead of saying "let's make different models for different conditions with the same causes (variables) in them" it says: "do not distinguish between these things as far as you can measure either causes or conditions". It makes analysis technically simpler, but does not undermine its quality. In fact, statisticians very often do not distinguish these (conditions like "weather", "days of week", etc., are typically considered just as variables) – so, I agree with this practice. And although it sounds weird to hear "Sunday is a cause of big sales", but for simplicity sake it's better to accept it, than try to create different techniques for these things.

4. One a different note, consider this: if it is known that direct cause of an illness is a particular type of bacteria, but these bacteria have different (known or not) probability to live in males and females. A gender as a "cause-condition combination" could be legitimately considered as a cause of the illness. As earlier, to say that gender is a "cause of the illness" is to simplify the language. It is a cause just in a sense, that physically women and men are "charged differently" with certain level of "direct causal force (bacteria)", which needs to be determined by the model. What is important – the fact that something (even Sundays) create specific yields for the outcome of interest to be detected. But the logic that causes are always some kind of "forces" should not be forgotten (see formal analysis of forces and fields in social life in [8]). It helps to distinguish between two situations considered in 4.

5. Example b) above describes situation, where there are quite obvious two types of causal inference – one for man and another for women. But examples a) and especially c) illustrate something different. Yes, one may assume that high income is a "cause" to buy an expensive care. But what about the "others, i.e. when X=0 in our notation? Do they also "cause" some level of Lombarghini sales, much smaller than for subgroup X=1 or we may treat purchases in that group as "random", i.e. it happened, but not because they belong to X=0 (it would be causal then), but just because "sometimes purchases happen regardless of income"? This is actually one of the key questions in causal analysis, but as far as I know it is not specifically understood and recognized.

The distinction should lay in a presence or absence of "charging force". There are variables where each value is "charged" with potential to create an outcome (as in "man – women" example), and other, where just one value does that in meaningful way, like "other income". Or, say, "Spanish" descent can produce some cultural traits, i.e. can force to move something, but "Non-Spanish" cannot. These two variables will be coded as binary, but they should be treated differently from causal point of view (not like in regression analysis or statistical learning). Another example: a variable "Income" with grade "Income from 50 to 75 thousand dollars" (1) and all others (i.e. smaller and higher) as 0. It is hard to expect that 1 is methodologically equal to 0 and one should look for "yield" for such a grade, which combines both low and high income. But very many datasets, especially in marketing, are designed in that way.

Let's summarize main assumptions made so far, in concise form.

1. Cause is some ***force***, either physical, psychological or social, which drives the outcome (effect). It pushes events in certain directions that results in outcome.
2. Subject of the analysis are only ***singular outcomes***, like "causes of the wounding", not the network of causal relationships (like what are the reasons, why one wounded the other, etc., as in SEM framework).
3. Only "***direct causes***" are considered, i.e. forces, which drive the outcome, not forces, driving these direct forces.
4. ***Conditions of environment*** are non-distinguishable from "direct forces". As far as certain combination "direct cause – condition" is assumed to generate some outcome, it is considered at par with "direct cause".
5. Causes can be ***either sufficient or necessary*** for outcome to appear, it doesn't affect the modeling. The only thing which is known for sure – the outcome had some causes, but there is no way to say the opposite (that cause produced the outcome), unless it is specifically proven.
6. Each value of the potentially causal variable produces outcome with its own ***intrinsic probability***. This probability is an aggregated representation of many individual trees lying behind the scenes in each particular individual, which are cut by value of this variable.
7. Usually, there are some causes, which cannot be associated with any measured variables, but still produce outcomes – let's call then ***random causes***.
8. The purpose of causal analysis is to estimate the intrinsic probabilities (or "***coefficients of generation of outcomes***"), using observed data.

All that is a radical departure from regression-like (or, wider, statistical learning) models, as well as from traditional models of causal analysis. This model is not oriented on mutual variation of outcome and causes, but directly on internal relationship between "producer and the product". Particularly, if correlation between dependent and independent variables are zeroes – regression (and based on it SEM), Bayesian networks models, etc. would yields zero coefficients, i.e. whatever the model, causes are not to be captured. But in the proposed approach it is not what will happen.

2.1.Causal analysis model

Let assume for simplicity sake that all causal variables are nominal, and they also just originate qualitative output. Then one may say, that each grade of the causal variable has its probability of generating output, regardless of values of the other variables. Random cause has the same effect.

Technically, if there is just one causal variable X with values Xo and X1 and binary outcome variable Y, the mechanism of effect generation will be like that:

$$\mathrm{Y} = \begin{cases} 1, \mathrm{if}(\mathrm{Y} \in \mathrm{A}(\mathrm{X}|1)\mathrm{or}\mathrm{Y} \in \mathrm{A}(\mathrm{X}|0)\mathrm{or}\mathrm{Y} \in \mathrm{A}(\mathrm{R})) \\ \qquad\qquad 0 \text{ otherwise} \end{cases} \qquad (1)$$

where $A(*)$ is a set of ones generated by either source, R stands for the "random generator".

From (1) follows that condition for Y=1 to occur is an “or” (disjunction, or union) function: values from two different sources –associated with measured variables and random – are not to be summed up to calculate total number of effect occurrence. The same value Y=1 would appear when either one or several causes would force it to appear.

It means that the total probability of the outcome would follow the rule of sum of probabilities, because random causes and “variables determined causes” are independent of each other. If one assumes, that both Xo and X1 generate outcomes Y=1with intrinsic probabilities $1 \geq \alpha \geq 0$ and $1 \geq \beta \geq 0$, while probability of the random outcome is ρ, the adequate equations are:

$$S_0 = \alpha + \rho - \alpha\rho \qquad (2)$$

$$S_1 = \beta + \rho - \beta\rho,$$

where S_0 and $S_1$are total probabilities of Y=1 in groups Xo and X1 respectively (the only measurable directly quantity).

For two causal variables there are four zones of intersection, and in each works the same rule:

$$\begin{aligned}
S_{11} &= \alpha_1 + \beta_1 + \rho - \alpha_1\beta_1 - \alpha_1\rho - \beta_1\rho + \alpha_1\beta_1\rho \\
S_{12} &= \alpha_1 + \beta_2 + \rho - \alpha_1\beta_2 - \alpha_1\rho - \beta_2\rho + \alpha_1\beta_2\rho \\
S_{21} &= \alpha_2 + \beta_1 + \rho - \alpha_2\beta_1 - \alpha_2\rho - \beta_1\rho + \alpha_2\beta_1\rho \qquad (3) \\
S_{22} &= \alpha_2 + \beta_2 + \rho - \alpha_2\beta_2 - \alpha_2\rho - \beta_2\rho + \alpha_2\beta_2\rho
\end{aligned}$$

where $\alpha_1$and $\beta_1$are the probabilities (frequencies) of grades of different variables.

If number of binary causal variables is K, then number of unknown values like α, β, ρ is 2K+1, but number of nonlinear equations like (2) and (3) ~O(2^K), i.e. started from K=3 number of equations exceeds number of variables and system became over determined.

The systems of equations like (3) for any *K*, according to Dr. B. Goldengorin (private conversation, for which I’m thankful to him), could be reduced to integer non-linear optimization problem, to be solved by so called Boolean, pseudo-Boolean approach [9]. It is possible, that some other methods could be applied.

The mathematical problems of how to solve these equations, how to make generalizations to non-binary causal and outcome variables and so on are not topic of this article. The simple heuristics I used to find the coefficients found perfect solutions on generated data, where regression estimates were deadly wrong. It is not surprising: this model does not do “summation of the causes” (which no one knows what is that about), does not create complex equations of the unknown nature, but modestly tries to separate different reasons one from another. In particular, it begs for completely different view to the multicollinearity problem, because intersection of the X variables just means that more outcomes from two or more sources need to be separated – moreover, it is built to work with collinear variables. This and many other interesting questions should be addressed in a future.

References

1. S. Lipovetsky (2013) Data Fusion in Several Algorithms. Presentation on JSM 2013, published in this Proceedings

2. I. Mandel (2011) Demography, Geography and Marketing: Telmar Centab – the largest US Census based study, Proceedings of the Joint Statistical Meeting, The American Statistical Association, Miami, FL, 2691-2698.

3. J. Pearl, "Causal inference in statistics: An overview," Statistics Surveys, 3:96--146, 2009

4. Eric D. Beinhocker The Origin of Wealth: The Radical Remaking of Economics and What it Means for Business and Society, Harvard Business Review Press, 2007

5. C. P. Águeda. Investigación Causality in Science (2011) Revista "Pensamiento Matemático" – N. 1. (see Zadeh's text in 6

6. I. Mandel Sociosystemics, statistics, decisions. Model Assisted Statistics and Applications 6 (2011) 163–217

7. F. Russoa, G. Wunschb and and M. Mouchartc (2012) Potential Outcomes, Counterfactuals and Structural Modelling. Causal Approaches in the Social Sciences. http://www.lofs.ucl.ac.be/fisc/staff/russo/files_talks/Counterf.-beamer-present--080707.pdf

8. D. Kuznetsov and I. Mandel, Statistical Physics of Media Processes: Mediaphysics. Physica A 377 (2007), 253–268.

9. B. Goldengorin, D. Krushinsky, P. M. Pardalos. Cell Formation in Industrial Engineering: Theory, Algorithms and Experiments. Springer, USA,2013, 218 pp.