PROBABILISTIC CAUSATION

An account of the effectiveness of graph-based probability theory for causal analyses

Dissertation

MSc Data Science

School of Mathematics and Physics
University of Sussex

By - Indranuj Banerjee
Student ID - 244853
Supervised by – Professor Enrico Scalas

Statement of Originality

This is to state that the following work titled "**PROBABILISTIC CAUSATION: An account of the effectiveness of graph-based probability theory for causal analyses**" is an original work by me and is being submitted for the fulfilment of the degree of MSc in Data Science at the University of Sussex. I declare that no part of the work has been submitted before for the degree of MSc in Data Science or any other degree or qualification at the University of Sussex or any other institution.

Indranuj Banerjee

Date - 01/09/2022

Acknowledgement

I would like to acknowledge my supervisor, professor Enrico Scalas for guiding me through the dissertation work. The resource materials, research papers and crash courses shared by him and discussions with him helped me grasp the concept of probabilistic causation with less difficulties. I thank professor Scalas for being considerate during the course of the dissertation work and not putting me under any kind of pressure or stress.

Abstract

When it comes to causation, there are two major schools of thought, The more traditional of which is deterministic. Whereas some believe that the universe is an uncertain place and most of it cannot be accurately known but can only be estimated till a certain degree of accuracy. This dissertation dives into the history of causation to understand its earlier notions and assess the various opinions of different theories. The report then goes on to describe probabilistic causation in terms of Bayesian network theory and conditional probability. At the end, it aims to analyze a specific scenario of football transfer to determine whether a football players performance causally affects his transfer fee.

Table of Contents

Table of figures

Introduction

"Shallow men believe in luck. Strong men believe in cause and effect"
-- Ralph Waldo Emerson

The concept of causality or that a cause leads to an effect is in the bedrock of human consciousness. It has its roots in philosophy; or *metaphysics* to be precise, which roughly translates to a branch of philosophy that deals with concepts that transcend the boundaries of known science or even religion by using logic and from a reasonable point of view. Given the age of human civilization, one might safely assume that pondering upon how a cause leads to an effect and what exactly defines a *cause* has been around for a while. Since ancient times, humans have been attributing natural phenomenon to arbitrary causative agents. Although they were not based on solid evidence or any evidence for that matter, the notion of something happening due to something causing it was well established. But the earliest formal theory of causality is accredited to Aristotle for his treatise *Metaphysics* where the term actually came into existence. The present-day discussions about causality, however, hold the work of David Hume with higher respect. Let us pause and think for ourselves about what a cause is. A succinct definition would be that a *cause* is any action performed or any phenomena occurring in time and space which directly and/or indirectly affects objects or circumstances at a later time in space which we call the *effect*. From a human sensory perspective, if two things are perceived to occur in conjunction (at different times) on multiple occasions with one succeeding the other every time, one might create an impression that the first thing causes the other. But there are many other events, the associations between which, are not realized or perceived naturally by our cognitive senses. This report delves to explore a brief history about the notion of causation and causal effects, its various schools of thoughts and lastly it explores associations that the human senses don't experience and asks if it is possible to develop models for these kinds of causal relations under empirical or observational settings. As it turns out, statistics and probability theory provide some robust tools that help us analyze such scenarios and estimate the causal connection between two events and answer counterfactual questions such as "how would it affect the health of a person if they follow diet B instead of diet A?" or "How would it affect sales if a restaurant opens only during peak hours of the day instead of staying open all day?". Questions such as these cannot be answered by the mind itself. We might be able to draw an association or a prior idea about how it might occur but there are certain things one needs to look out for. We will gradually course into that starting from the next section, i.e., a brief history of causation.

A brief history of causation

Causation in philosophy and Humean Causality

The human mind can make causal inferences in two directions. A physician might look at the symptoms and deduce what a patient is suffering from. A chemist knows about the nature of a solution from the colour of litmus paper. A person sitting inside a room hears thunder and can say that there must have been lightning. These are examples of **deductions** made from an effect about what might have brought it on, viz. the cause. On the other hand, a seismologist knows that if a volcano erupts, a shockwave will ensue. If milk is left out overnight, it will be bad in the morning. And you the reader, know by now, that these are examples of **predictions** of the effect by analysing initial conditions, based on what you read earlier; all based on one premise, "It always happens like that."

Scottish philosopher David Hume in his work*, A Treatise of Human Nature* [1]*,* proposed that all thoughts or perceptions of the human mind are based on experiences, and recurrence of the experiences cause us to form a belief about the world by what he called the *copy principle* [2]. He claims that the human mind forms causal inferences in the same way; only after repetitive experience of events happening in conjunction. From collective experience (the irony) we can all agree on this. Ever since childhood, our inquisitive and curious minds make us question everything as we watch them unfold in front of us. The human consciousness does capture the essence of causation by perception of changes in a system as it moves forward through time. But defining causation is difficult. It is a concept whose real purpose or notion is unknow but using which we describe other phenomena in the universe. This could mean causation is a part of the foundational construct of reality that evades our understanding, at least for now.

The theory proposed by Hume was a rather sceptical view on causality for his time. It attempted to explain what a *cause* actually means along with how the human mind draws inferences from causal relations from observations which may be affected by subjective beliefs [1] [2]. He says that the notion of causal relation between two events, e.g., lightning strike followed by thunder, or a billiards ball rolling and hitting another ball that causes it to move, is only because we experience both the events in conjunction on every instance. Hume claims that nothing about the nature of the events tells us how one causes the other as in we do not perceive the process that underlies the causation. All we are left with is a mental impression of the state of the events or objects take partake in the causal relation. Nothing about the lightning strike tells us how and why the sound of thunder will follow. Nothing about the feel or the shape of a billiards ball tells us how it makes another ball move if it hits. We only see one ball move, hit another ball and stop, while the other starts moving from a stationary state. And with repeated observations, we assume that since both happen together, they might be causally connected. Repeated observation of such kinds changes our expectation about the occurrence of an event as in when we experience some event of such kind, we expect the next event to follow. To this David Hume says that the *necessity* for causation, therefore, need not be embedded in nature but can only be in the mind [1] [2] [3].

Humean causality [1] defines a causal relation as follows:

"An object precedent and contiguous to another, and where all the objects resembling the former are placed in like relations of precedency and contiguity to those objects that resemble the latter."

This definition is actually a set of conditions or limits to causality that, Hume claims, help distinguish "true" causal relations from non-causal relations [4], which are:

- The cause and effect must be contiguous in space-time.
- The cause must always precede the effect in time.
- The cause and effect must be of the same type. For example, if C causes E, then every time a type C event occurs, type E event has to occur and in exactly the same way as before.

If a pair of repetitive perceived observations satisfies the above conditions, then it can be said that the first event caused the other. The first point limits causality, in terms of sensory perceptions, to close distances where the cause and effect are connected directly or through intermediate chain of events. The second point emphasises on the chronology of the events and states that causation is asymmetric. The third condition asks for a like relation between cause and effect on every instance. Therefore, Humean causality is said to be a regularity theory [4]. It only establishes a relation between the cause and effect but does not signify any causal transmission between the two. Hume states that observation of the cause is sufficient to expect the effect. It does not actually necessitate the effect. There is only constant conjunction and with repeated experience, our mind starts to draw inferences even about future events in the chains of regularity that have not occurred yet. David Hume's causality of the mind is equivalent to causation in the world if there is a hypothetical omniscient being who perceives all events and can infer all causal relations from experience by the copy principle.

German philosopher Immanuel Kant later on built on some of the ideas of Hume wherein he split the workings of the human mind in two parts [5] called *sensibility,* which is the information one gets from experience and *the understanding,* which is the various ways the mind processes the information. Kant differed from Hume on the belief of causal necessity pertaining to natural laws. Hume's treatise did not mandate this but Kant argued that laws of nature should be a necessary condition for causality as they are causal laws. In fact, if we follow the definition of a causal relation to be a contiguous chain of cause and effect, then everything is essentially either a cause or an effect of something in a vast causal network. Kant even goes further to claim that since associations are based on subjective experiences and beliefs of the mind, it would mean *our* perception of some event is what dictates *us* to write the laws of nature the way we do to represent a certain phenomenon [5] . One could assume the true nature of those laws, outside of our perception to be something else altogether. For example, all the laws of physics take for granted as a matter of *fact* that time is unidirectional. It may stretch or warp but it always moves from before to after. and only events from the past can affect the present or the future. A new idea of retro causality has been proposed to explain quantum entanglement that is based on the premise that a subatomic particle can run back in time to influence the other.

Causal Determinism

Philosophers like Kant and Hume's view on causality is atomic (deterministic) and the traditional way of thinking about causality. Aristotle and the stoics of ancient Greece believed that everything is fated. Not in a general sense, but with respect to what happens just before it. In the contiguous chain of cause – effect pairs, if we pick any random pair that are directly causally related, the cause (in this case, the event that occurred first) is inevitable and as a result the next event has to follow, thereby making the latter inevitable as well and so on. With the popularity of classical mechanics, majorly due to the works of Sir Isaac Newton, the idea that the universe is deterministic, became the common notion for a long time. Famous mathematicians such as Pierre-Simon Laplace and Albert Einstein were both strong proponents of determinism. In 1814, In one of his publications about causal determinism titled *A philosophical essay on probabilities*, Laplace talks about a hypothetical creature (popularly known as Laplace's demon). One who knows the location and momentum of every atom in the universe, and who can calculate the location and momentum of any particle at any time in the past or the future using classical mechanics. The vision of causal determinism therefore denies the existence of free will as it believes everything is predestined to a certain fate and if the fate is inscribed in a law of nature, it is possible to predict it. This is called strict determinism. But almost a century later, with the advent of quantum mechanics, we got to know that it is not the case after all. As Werner Heisenberg stated in the famous uncertainty principle that it is not possible to predict the value of a quantity (position or momentum) with certainty, even if all initial conditions are specified. Classical physics describes our macroscopic world to a high degree of certainty. This is the reason why bridges do not crack during the summer heat due expanding parts because engineers can predict the expansion and know exactly the right amount of gap to leave between those parts to allow for the expansion. Also, reason why scientists are able to successfully launch satellites and rockets into space. Another

example is the neural network in our brains, that work reliably transmitting signals to various parts of our body to keep them functioning. Therefore, in a practical macroscopic world, there is no reason to believe that events (causal events) could be anything but deterministic. [6] and [7] make critical reviews of probabilistic causation in terms of determinism. The author refers to the practice of attributing high probability values to signify a causal relation and terms it crypto determinism, stating that causality is deterministic when it comes to study of individual macroscopic objects but requires statistical framework to accommodate studies based on population where it is difficult to keep track of all causes and effects.

But still there are few issues that regularity or strictly deterministic theories cannot address, such as [8]:

- Imperfect regularities – When the cause is not invariably followed by its effect. As is the case in studying how a certain virus affects the human body, for example. When exposed to a certain virus, only some portion of the population show symptoms of the disease whereas other do not. When an event is not known to occur, no other event can be called its sufficient cause.
- Irrelevance – Probabilistic theories ask that the cause make a difference to the effect. For example, a drop of water will evaporate irrespective of if it is put on a warm or cool plate. The state of the plate makes no difference.
- Spurious correlations – Sometimes, an unobserved event is the cause for two separate events. In an experimental setup, the latter are observed to be correlated and might lead one to believe one to be the cause of another or lead to wrong analyses, as variables of interest might be causally dependent on the unobserved variable. Probabilistic theories provide means to eliminate such issues efficiently.

With the rise and boom of quantum mechanics in the scientific sector, rose the idea that causality could be probabilistic because the nature of the building blocks of reality themselves were shown to be probabilistic [8]. This is due to the measurement problem in quantum mechanics wherein a particle if not measured (some theorists translate 'measured' to 'perceived by the human senses') is a superimposition of several states described by their wave function. The wave function assigns *amplitudes* to the different measurement outcomes of the particle with the probability of finding the particle with the measurement is its amplitude squared. Einstein couldn't reconcile himself with quantum mechanics and is famously know for saying, "God does not play dice with the universe." And we also stated before that the macroscopic world governed by Newtonian mechanics is deterministic with all its laws. So where does probabilistic view fit in?

Probability theory

The concept of probability or chance is not a new concept by far, the origins of which can be traced back to the period of Renaissance. It was considered a tool for the low sciences such as alchemy or medicine [9] as it involved the study of opinions or verbal statements. Whereas the high sciences were astronomy and mechanics that required mathematical and empirical knowledge. The modern-day origin of probability dates back to the 17th century and is attributed to a correspondence between Pierre de Fermat and Blaise Pascal over a gambling game [10]. Later on in the 18th century, Jakob Bernoulli in an attempt to reason causes from their effects, presented his law of large numbers where he claimed that the proportion of number of heads in a coin toss and the proportion of male births will converge to a value p as the number of observations increase. He said that when the value of p is not known, it can be inferred by someone with greater experience [10] [11].

One of the very first accounts of using probability theory in the study of social behaviour was when Laplace wanted probability to serve as the basis for the moral sciences [10], along with several of his more politically active mathematicians, most notably Marquis de Condorcet. This primarily took the shape of judicial and electoral probability, addressing some of the main issues raised by the philosophers and critics of the Enlightenment. For the French mathematicians, elections and justice had formal similarities. The capacity to increase the probability that a jury or voter would reach the right decision was a key issue in each case. One of

the components included testimony, a well-known area of probability theory. Condorcet developed a sustained interest in the testimony given during trials, proposing to evaluate the veracity of each witness's testimony by taking into account the percentage of times he had previously stated the truth before combining the testimony of multiple witnesses using inverse probabilities.

Although being propagated by numerous mathematicians, the theory of probability was put under a lot of scrutiny mostly because of the sceptical problem of generalizations. [9] Is it ever possible to believe a generalisation from a small sample of observed cases without conducting a comprehensive survey? It may be closely tied to the Humean causality theory, which claims that no collection of prior observations provides sufficient evidence to predict if an event will play out in the same way in the future as in the past [2]. Our expectations lack justification and are generated by mere habit and custom. The scepticism about statistical studies in earlier times can be corroborated by Greek philosopher Sextus Empiricus' quote from his book Outlines of Pyrrhonism;

"It is also easy, I consider, to set aside the method of induction. For, when they propose to establish the universal from the particulars by means of induction, they will effect this by a review either of all or of some of the particular instances. But if they review some, the induction will be insecure, since some of the particulars omitted in the induction may contravene the universal; while if they are to review all, they will be toiling at the impossible, since the particulars are infinite and indefinite. Thus, on both grounds, as I think, the consequence is that induction is invalidated."

But still, people could differentiate a good induction from a bad one and therefore the task was to classify the degrees of observations used as evidence to support the said induction. Probabilistic studies of causation were significantly advanced in the 20th century by statisticians with an inclination towards philosophy. Determinists maintain that causality is in fact deterministic and the fact that there are imperfect regularities (stated above) and have to be studied from a probabilistic point of view is interpreted as an illusion of determinism. Because, an induction about a causal relation using probabilistic approach is made only when observations of the event occurring when the cause is present is high enough, i.e., P(Effect |Cause) is high. Sceptics [6] claim that determinism is specific case of the more generalized probability approach where P(Effect |Cause) = 1.

The probabilistic approach to understand causation and trace regularities in the social structure has been successful as social structures involve randomness to a high degree and statistics has been found to provide good inferences about a population given that adequate observations are made with proper representation of population (unbiased population). On top of that, the statistical language allows us to accommodate for the margin of error in our calculations so that any assertion made with sound logic and solid observations can be rarely falsified.

Mental Model Theory (MMT)

The theory of mental models provides a deterministic account of the everyday meaning of causation. C causes E is a statement that can be decomposed into 3 possibilities (in causal analyses dealing with opinions or verbal assertions, cause is generally denoted by C and effect by E), namely:

I.	C	then	E
II.	Not-C	then	E
III.	Not-C	then	not-E

With the condition that C always precedes E in time. If C and E have been observed to occur, then (I) is a fact and (II) & (III) are counterfactual possibilities which support assertions such as *'If C had not happened then E wouldn't happen.'* Conversely, if C, the cause, is not observed to occur, the counterfactual changes to the form *'If C had happened then E would happen.'* [12]

MMT makes distinction between a cause and an enabling condition. It considers an enabling condition as a constant factor that does not violate the presupposed idea of determinism and a cause as non-constant with respect to the variables in question outside the natural progression of events [13]. For example, in the causal analysis of the start of a wild fire, as per MMT, the presence of oxygen is considered an enabling condition and not a cause as it is constant for everywhere. Whereas friction between dry branches or a rogue spark is treated as a cause because it is not naturally expected to be present and has to be introduced by some way into the natural progression of events. But even in this example, one has to ask the question, that although the final product of the forest fire arising from a spark igniting dead branches is deterministic. But even then, there are certain conditions, like the weather or landscape that have to be satisfied for the fire to build.

The concept of causation in the MMT makes no assumptions about whether or not every event has a cause or whether events can start causal chains. However, it's frequently claimed that interventions (manually assigning a certain value to a variable out of normal causal chain) have their own unique rationale. For instance, if one were to notice that someone isn't obese, they would assume that the individual doesn't overeat given the causal claim that eating too much leads to obesity. But let's say the person's medication for preventing weight is discovered. They could no longer conclude that the lack of obesity meant that overeating was not to blame. The medication neutralises the consequences of overeating. A basic understanding of the following premises is all that is required.

- Overeating causes obesity.
- Taking an anti-obesity pill prevents obesity.

And to understand that the second premise is more important than the first.

In MMT it is considered that a single false observation is enough to refute an inference as false [12] [13]. And to understand whether everyday causation is deterministic or probabilistic, [13] presented individuals with certain assertions and asked them the number of contradictions required by each of them to refute an assertion. According to researchers, people are prone to two biases: Supposing that they are given two events A and B and asked if there is a causal relation between them, they focus on instances in which A and B co-occur to infer that yes, causal relation is there, which corresponds to the mental models of the concept; and they take a causal relation to mean that A is sufficient for B, so the relation is refuted by the occurrence of A without B. It's unclear their account includes enabling assumptions. Most subjects were found to dismiss causal claims after a single counterfactual observation on average giving the idea that everyday causation is perceived as deterministic. [14] provides a critique of [13] and states that dependency and production are two essential frameworks that can be used to group the many notions. In the notion of dependence, if event A hadn't happened, event B wouldn't have happened, then event A causes event B. Causation in the sense of production is dependent on the idea of force, power, or mechanism. Most ideas in psychology can be linked to or drawn from philosophical concepts, including the force dynamics model, probabilistic contrast models, and the mechanistic explanation of causation which support the coexistence of various representations in causal reasoning on the grounds that "no one concept of causality should be able to account for the complete variety of occurrences involved with causation". [14] also talks about a new theory of causal pluralism, according to which, "humans use numerous, mutually-interactive representations that can be grounded in distinct causality frameworks simultaneously in daily causal reasoning". It then goes on to provide some logical drawbacks of the MMT experiment as follows:

I. Their tasks presupposed the refutation of causal assertions on the basis of individually isolated and distinct options rather than a probabilistic criterion. Then, they used a single refutation to demonstrate the deterministic concept of causation. In an extreme circumstance, this logic is similar to prematurely pronouncing someone as guilty, then asking what case can be framed to falsify their innocence, and using the conjured-up case to pass their sentence.
II. It is insufficient for a single refutation to be the criterion of determinism. A refutable assertion may be neither necessarily true nor probably true, but only actually true.

III. It is unnecessary for a single refutation to be the criterion of determinism. A real strict causal law cannot be refuted by a single observation. This is because the causal law itself is more robust than an accidental observation of a counterexample.

To remedy the above flaws, [14] conducted similar experiments of their own by subjecting individuals to more generalized assertions rather than definite statements across different domains and taking their stance on the causal relations in those assertions. Their results displayed that the domain of the assertion influences the deterministic or probabilistic nature of the causation.

"In the four domains we tested, the physical and physiological domains have a robust tendency to yield deterministic interpretations, but with a significant difference, in that the former is stronger than the latter. In contrast, the psychological and socio-economic domains have a robust tendency to yield probabilistic interpretations, but with a significant difference in that the former is weaker than the latter. This demonstrates a pluralistic rather than a monistic perspective of the modal conception of causation [14]".

While it is true that the probabilistic nature of the quantum world does not affect the deterministic nature of the macroscopic world, but we regularly find randomness associated with the state of the effects that have a presupposed deterministic notion about them. For example, the rolling of a die is deterministic. With our perception of dice rolling that we have gathered from experience and knowing that it is us who are rolling the die, we start the causal chain. The die rolls and stops at a point which we can know deterministically using Newtonian mechanics but the number that comes up on the die is completely probabilistic in nature. If this was a game and the next move depended upon the number coming up on the die, it would eventually lead to different outcomes. Of course, these outcomes can all be classified under the same category, that "throwing a die will make it roll and then stop". Or "if the game is played, a participant will win". But these cannot be the realizations of causation. A closer analysis of causation requires it to be probabilistic in nature as most real-world systems contain randomness due to unobserved background conditions and on certain levels or a certain combination of covariates, there are regularities. In essence, deterministic systems can be fit into the notion of probability by citing them as special cases as stated before.

Because of this, it is possible to properly analyse causal relationships by being aware of the variables that are at play and, in the case of multiple variables, how they interact to produce a particular outcome (the effect). In order to deal with the consequences of the unseen background, appropriate information is also necessary. We can infer causal linkages from observations using the powerful statistical and graphical tools provided by the theory of probabilistic causation.

Probabilistic causation and its literary review

The laws of proportions that control repetitive trials, such as the results of gambling machines, seem to have no compelling reason to combine with beliefs, which are mental dispositions concerning unpredictable and frequently unobservable events. Probability theory's ability to articulate practical qualitative correlations between beliefs and to process these relationships in a way that results in intuitively believable conclusions is its main appeal, at least in situations where intuitive judgments are strong. The following is a quote from Judea Pearl's book [15] *Probabilistic reasoning in Intelligent systems*:

"*What we wish to stress here is that the fortunate match between human intuition and the laws of proportions is not a coincidence. It came about because beliefs are formed not in a vacuum but rather as a distillation of sensory experiences. For reasons of storage economy and generality we forget the actual experiences and retain their mental impressions in the forms of averages, weights, or (more vividly) abstract qualitative relationships that help us determine future actions.*"

After quantum physics shook the faith in determinism [8] and with advancements in statistical tools, such as nonparametric structural equations, graphical models and a symbiosis between counterfactual and graphical models [16], it was possible to study causality from a probabilistic perspective. Most empirical studies in health, social and behavioural sciences are causal as they deal with deriving inferences about effectiveness of a certain treatment on a population, selection bias in a candidate selection process or predicting outcomes given a different set of conditions. It should come as no surprise that at the core of these investigations is to clarify cause-and-effect relationships among the variables in the respective domain. And therefore, has come under speculation from sceptics about the veracity of the assumptions made to elucidate the cause-effect relation along with the data collection methods. Even very recently, there was scepticism toward probabilistic studies since when McCarthy and Hayes declared probabilities to be "epistemologically deficient" in 1969, artificial intelligence researchers resolutely avoided probability [15].

Regression and other estimation approaches serve as examples of standard statistical analysis, which aims to infer a distribution's parameters from samples taken from it. With the help of such parameters, one can infer relationships between variables, calculate the probabilities of events in the past and the future, and update the probabilities of events in response to fresh data or improved measurements. So long as the experimental settings don't change, normal statistical analysis does a good job of handling these jobs. Causal analysis on the other hand aims to infer likelihoods of events upon changing experimental conditions which might come about naturally (such as weather conditions) or be caused by an intervention (such as applying a treatment), identifying causes of reported events, and assessing responsibility and attribution (i.e., if A was necessary to cause B). Any correlation between two variables under static conditions is merely an association and it cannot be assumed that there is a causal relation between them. To substantiate a causal relation, the association must hold between the two variables even across changing conditions, accompanied by an assertion under some premise that justifies the existence of a causal relation between the two variables. Standard statistical methods only provide us with the distribution of a random variable which is analogous to the 3-dimensional representation of a geometric object. It helps us visualize the object (realize the behaviour of the random variable) from all angles but does not give any information on how the object will change if manipulated. The information is encoded in the causal assumptions in the form of graphs or structural equations. The two main mottos of probabilistic causation are (i.) Causes raise the probability of an effect [8] and (ii.) Correlation does not mean causation [16]. Causation cannot be identified by the joint probability distribution alone, there must be some logic behind every claim of causal relations.

Associations might also represent the presence of a confounding variable. Confounding deals with the discrepancy in the measured association between two variables in an observation and under ideal conditions. It

is possible to provide straightforward instances that demonstrate how the associational requirement is not required nor sufficient. In other words, certain confounders might not be linked with either X or Y, while other non-confounders might be associated with both X and Y. This further suggests that confounding bias cannot be identified or addressed using statistical approaches alone, not even the most advanced methods that claim to "control for confounders," like stepwise selection or collapsibility-based methods. But, as we will see later on, methods such as graphical models and structural equation models make visualizing and adjusting for the confounding variables much easier. Before an adjustment may properly compensate for confounding bias, one needs make certain assumptions about the causal links in the problem, specifically about how the possible "confounders" affect other variables in the problem. The highly emphasized distinction between association and causation also calls for different symbols and notations to represent causal relation analyses to differentiate them from ordinary statistical analyses. [16] claims that 2 main obstacles for the acceptance of causal analysis in the world of statistics are scepticism towards the nature of causal assumptions and the need for a different notation for representing causal relations. Judea Pearl provides a notation and calculus for analysing causal relations [17]. Further papers [15] [16] [18] and this blog [19] show the implications and applications of the notation in further detail.

Primarily, the causal relations studied are of two types, general or type-level causations and actual or token-level causation. The former pertains to a general assertion, such as “eating lots of chocolate causes dental cavities” whereas the latter points to more specific events like “His tendency to get agitated quickly is because of a strained relation with his father”. A similar contrast is that actual causation is concerned with how events really transpire in a particular situation, as opposed to general causation, which is concerned with the full range of potential outcomes [8].

Concept of probability raising

The central idea of probabilistic causation is that causes raise the probability of an effect [8] [20], coupled with the chronological constraint that the cause always precedes the effect in time. If we consider a cause C and an effect E, then as per the above statement, we can say

$$P(E|C) > P(E|{\sim}C)$$

Which means that the probability that E occurs when C occurs is greater than that when C does not occur. We can say that C is a potential cause of E if the above inequality holds. Assertions made using the inequality can accommodate imperfect regularities, i.e., the condition will still hold even if the cause is not invariably followed by the effect. The condition may even be the exact opposite, as in $P(E|C) < P(E|{\sim}C)$ in which case it would mean that C is the prohibitor of E, but it can never be equal. This form of representation does not represent the asymmetrical nature of causality nor does it deal with spurious correlations or confounders.

The Direction of Time by Hans Reichenbach [21] was released posthumously in 1956. Its focus is on the causes of temporally asymmetric occurrences, especially the rise in entropy required by the second law of thermodynamics. The first completely developed probabilistic theory of causation is presented by him in this work. In it, Reichenbach introduced the term *screening-off* to describe the following kind relationship where;

$P\left(E \,\middle|\, A\&C\right) = P(E|C)$; the expression translates to C screening off A from E. If P(A & C) > 0 then;

$P(A\ \&\ E|C) = P(A|C) \times P(E|C)$ which means that A and E are independent conditional on C.

There are two ways Reichenbach identified that this could happen. (i.) If C is the intermediate variable between A & E represented by A → C → E where the → denotes causal direction. This means that A causes C which in turn causes E. For example, Ozone (A) can become water (E) only by decomposing into oxygen (C) first. But no oxygen molecule that came from ozone are more likely to form water than any other oxygen molecule that came from different sources. Hence the occurrence of E is independent of A conditional on C.

The other way that this can happen is if C is a common cause of both A and E; denoted by A ← C → E. From the diagram it can be inferred that the occurrence of A is completely independent of E although both are dependent on C. If A in turn has causal relation with E, then C becomes what is known as a confounder. Reichenbach used the concept of screening off to tackle the problem of confounding. For example, the drop in atmospheric pressure causes the mercury level in a barometer to drop. At the same time, it might also cause a storm. Since the event of the atmospheric pressure dropping is not apparent to human sensation, a correlation may be drawn between the mercury level falling and a storm happening. In Reichenbach's words, C_{t1} is a cause of E_{t2} if and only if:

- t2 > t1; where t1 & t2 denote time
- $P(E_{t2} | C_{t1}) > P(E_{t2} | \sim C_{t1})$
- There is no other event X at a time t3 ≤ t1 that might screen off E_{t2} from C_{t1}.

Reichenbach tackled the issue of confounding variables by his formulation of the common cause principle. Following from our previous example that C causes both A and E and they show a positive correlation such that P(A & E) > P(A) x P(E), but assuming that we know that neither A nor E are causes of each other. Then Reichenbach states that the following relations will hold with respect to the common cause C:

I. $0 < P(C) < 1$
II. $P(A \,\&\, E|C) = P(A|C) \times P(E|C)$
III. $P(A \,\&\, E|\sim C) = P(A|\sim C) \times P(E|\sim C)$
IV. $P(A|C) > P(A|\sim C)$
V. $P(E|C) > P(E|\sim C)$

As we have seen before, this expression (II.) means that A and E are independent of each other when conditioned on C.

Simpson's paradox

Following from the common cause principle, we can see that the traditional point of view of probabilistic causation that causes raise the probability of effect is a necessary but not sufficient condition to signify causation. It may turn out positively correlated events are not causally related at all. [8] gives the following example. Suppose that smoking is strongly associated with living in a rural area. People who reside in rural areas are also significantly more likely to smoke. Lung cancer is brought on by smoking, but what if city pollution was a bigger contributor? Then it's possible that smokers have a lower overall risk of developing lung cancer than non-smokers; which we can represent as P(E|C) < P(E|~C). However, if we condition this expression on the staying in the country (F) or in the city (~F) then the expression can be extended to

- $P(E \,|\, C \,\&\, F) > P(E \,|\, \sim C \,\&\, F)$
- $P(E|C \,\&\, \sim F) < P(E|\sim C \,\&\, \sim F)$

These probabilistic inequality reversals are examples of "Simpson's Paradox." [20] Both Nancy Cartwright [22] and Brian Skyrms [23] noted the issue that Simpson's paradox poses for probabilistic theories of causality. The rectification proposed to this problem was to expand F to include all kinds of background conditions. A background condition is the combination of several factors. Such a confluence of factors is said to be "kept fixed". Therefore, we must declare what variables are to be held constant in order to specify what the background conditions will be. In our example this variable is F and the background condition is living in the country (F) or not living in the country (~F). Analysing the relation at every level of the variable F is called conditioning upon or stratification upon. But not all seemingly relevant background conditions can be conditioned upon by mere observation and require systematic selection. It is important to keep in mind that when building background

contexts for C and E, one should hold fixed not only positive causes of E that are independent of C, but also negative and mixed causes of E. In other words, one should hold fixed all factors that are causally relevant for E, with the exception of those which are causally connected to C [20].

However, causal statements are rarely seen in the mathematics of the majority of statistical literature. They only come into play when investigators verbally interpret particular associations or when they use verbal descriptions to support their presumptions. Despite its success, the mathematics of causal analysis remains mysterious to the majority of rank-and-file researchers, and the statistics-based sciences continue to greatly underuse its potential. The next part talks about the various mathematical ways causal analyses are performed.

Structural equation models and DAGs

Structural equation models (SEM) are mathematical representations of causal assertions. The credit of attempting to formulate such equations goes to Geneticist Wright. For example, if a random variable X denotes a disease and Y denotes symptoms of the disease, then [16] , [20] and [24] say the realizations of Y will be given as

$$y = \beta x + u$$

Where x – severity of the disease

y – severity of the symptoms.

u – other factors affecting Y apart from X

Notice the letters designating cause and effect (C and E) become X and Y, which is more associated with mathematical formulae involving input and output rather than verbal claims.

"*In interpreting this equation, one should think of a physical process whereby Nature examines the values of x and u and, accordingly, assigns to variable Y the value* $y = \beta x + u$*."* [16]

But it does not imply that $x = (y - u)/\beta$ like regular equations. To distinguish a structural equation representing a causal relation from ordinary equations, the former is always accompanied by a *causal path diagram* (as used in the section prob raising). The causal path diagram of the above relation can be written as X → Y. It is just a representation of the direction of causal influence and may be composed of a greater number of variables demonstrating how they influence each other through causal mechanisms. For example, our simple relation can be extended to

Figure 1: A causal path diagram capturing causal relationships among observed and unobserved factors.

Here we see two more variables V and U and they causally relate to X and Y respectively. V and U are called exogenous variables as they represent background conditions whereas X and Y are called endogenous as they are the model variables. As per SEM, the causal equations can be written as x = v and y = βx +u. The term β represents the path coefficient between X and Y which is the causal effect of X on Y. This roughly translates to the meaning that a unit change in X will result in β units change in Y [16] [24].

The graph representation of a causal network is advantageous in many aspects than the language of conditional probabilities. Firstly, it helps us picture the various relationships between the nodes and secondly, it allows us to systematically assess which nodes are relevant for the analysis. As stated by the pioneer of causal analysis using graphical models, Judea Pearl in [15],

"*Our goal is to make intentional systems operational by making relevance relationships explicit, thus curing the impotence of declarative statements such as P(B |A) = p. As mentioned earlier, the reason one cannot act on the basis of such declarations is that one must first make sure that other items in the knowledge base are irrelevant to B and hence can be ignored. The trick, therefore, is to encode knowledge in such a way that the ignorable is recognizable, or better yet, that the unignorable is quickly identified and is readily accessible.*"

Plus, the understanding of graphical representation spans across domains and any information regarding causal analyses can be more easily accessible for a vast array of professions like epidemiology, finance, genetics and computer science.

In causal graphs or **directed acyclic graphs (DAG)** [8] [17] [20] [24] as they are known:

- Each node represents a factor in the causal network while an arrow shows direct causal direction.
- Unobserved but known causal relations are denoted by a dashed arrow.
- Unobserved presumable common causes are denoted by a double headed dashed arrow.

In reading a path from a DAG, relational terms such as ancestor, descendant, parent and child are used whose meanings are self-explanatory. Three kinds of paths can be expected in a DAG

I. $X \rightarrow A \rightarrow Y$ Which signifies a causal translation. A translates the causal effect/force from X to Y.
II. $X \leftarrow A \rightarrow Y$ Which signifies confounding. A is a common cause of X and Y.
III. $X \rightarrow A \leftarrow Y$ Which signifies a collider. A blocks the propagation of causal effect from X to Y.

It's vital to remember that causal assumptions are contained in path diagrams not in the linkages themselves, but rather in the missing links. A missing arrow makes a firm commitment to a relationship with zero strength; an arrow just suggests the possibility of a causal connection, the strength of which is yet to be assessed [16] [18].

Types of questions pertaining to causality that can be analysed using graphical modelling may be figuring out the correct causal structure for a given set of variables, or to test the effects of interventions on a causal relation. This is one of the bigger aspects of a DAG that it allows us to answer question like "What is the probability that Y would take the value y if we were to intervene in the causal structure and set the value of X to x?". Can the controlled (post-intervention) distribution, P(Y = y |do(x)), be calculated from data governed by the pre-intervention distribution, P(z, x, y)? is the central query in the investigation of causal effects. This is the identification issue that causal analysts have focused a lot of effort on [20].

This identification would be possible whenever the model is Markovian, that is, the graph is acyclic (i.e., contains no directed cycles) and all the error components are jointly independent, according to a fundamental theorem in causal analysis. The following fundamental theorem can be used to define the conditions under which non-Markovian models [16] [20], such as those with correlated errors (coming from unmeasured confounders), permit identification.

The Markov Condition

In a causal DAG, often there are lots of variables that may or may not influence each other directly or indirectly. To ensure that the particular causal relation being analysed is not biased by the influence of other covariates or confounders so we have to make the desired relation conditionally independent of other covariates in the network [8]. Following the three Markov conditions ensure conditional independence. They are as follows:

For a graph G based on variable set V, the joint probability distribution P satisfies Markov condition if [8] [16] [20]

For every variable X in V, and every set of variables
$Y \subseteq V \setminus Desc(X), P(X \mid Par(X) \,\&\, Y) = P(X \mid Par(X))$ (MC screening off)

Let V = {X1, X2, …, Xn}. Then,
$P(X_1, X_2, \ldots, X_n) = \prod_i P(X_i \mid Par(X_i))$ (MC factorization)

Let X, Y ∈ V, Z ⊆ V \ {X, Y}. Then
$P(X, Y \mid Z) = P(X \mid Z) \times P(Y \mid Z)$ if Z d-separates X and Y in G. (MC d-separation)

Each of which are explained below. We can understand the probability statements above more intuitively if displayed in a graph (figures 2, 3, 4).

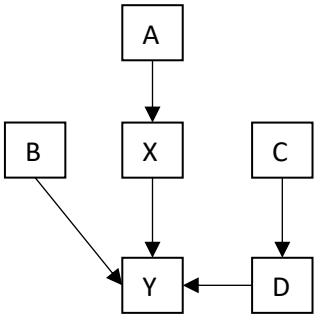


Figure 2: DAG to picture MC screening off

MC screening off – This states that the parents of a variable X screens off X from all other covariates in the causal network except its descendants. [16] [17]

In the adjacent graph, suppose we want to find the causal effect of X on Y. This has to be done by making sure that X is conditionally independent of any other covariates except its parents. The parent of X, i.e., A screens off X from B, C and D. We can see that the values of none of the three will affect the probability of what value X takes. The graph therefore satisfies MC screening off and causal relation between X → Y can be calculated.

MC factorization – This states that once we know the conditional probability distribution of each variable given its parents, we can compute the complete joint distribution over all of the variables [16] [17] [20].

The joint probability distribution for the adjacent graph is therefore given by:

$$P(A, B, C, D, X, Y) = P(A)P(B)P(C)P(X|A)P(D|C)P(Y|X, D)$$

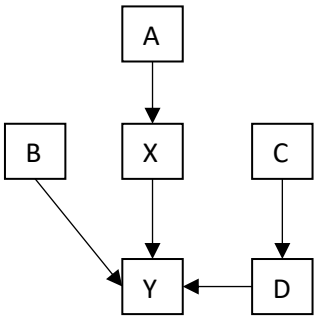


Figure 3: DAG to picture MC factorization

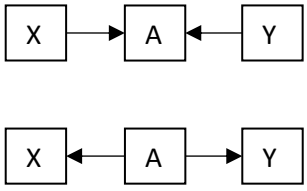


Figure 4: DAG to picture MC d-separation

MC d-separation – It is similar to screening off but uses a different logic.

If during path traversal through a DAG, causal relations as of the adjacent figure are found, in both the cases A d-separates X from Y. In the first case A is a collider and readily d-separates X from Y as it does not let any causal associations pass in any direction between X and Y making them independent of each other. On the other hand, the second kind of path shows a confounder. In this case, it is required to stratify by the confounding variable (A) so that X is independent of Y but conditional on A. This forms the basis of satisfying something called the *backdoor criterion* [16] [17].

In theory, the Markov conditions are only sufficient conditions for causal independence. It only shows conditional independence between variables in a causal network. For necessary conditions for probabilistic causality, one needs to refer to the minimality and faithfulness conditions.

The minimality condition states that the probability that an outcome variable takes on a certain value given a certain input is different from the probability that the outcome variable takes the same value with a different

input, given that all other background conditions are held constant [17]. Verbally, we can approach this concept as saying that something can be a cause of something if the chance of obtaining the latter changes if there is a change in the former. Mathematically, it is given by:

$$P(Y = y \mid X = x \,\&\, Z = z) \neq P(Y = y \mid X = x' \,\&\, Z = z) \qquad (1)$$

And the relative causative effect is given by

$$P(Y = y \mid X = x \,\&\, Z = z) - P(Y = y \mid X = x' \,\&\, Z = z) \qquad (2)$$

Or, even the ratio

$$\frac{P(Y = y \mid X = x \,\&\, Z = z)}{P(Y = y \mid X = x' \,\&\, Z = z)}$$

Interventionists make use of this idea to make comparative analyses and make inferences about events that may not have occurred. For example, what will be the probability of recovery on the whole if x = drug 1 is used and when x' = drug 2 is used on the entire population. If the drug has any effect on the recovery, then inequality 1 necessitates that and equation 2 helps us identify the better drug (one with higher probability of recovery). In which case the values X = x and X = x' are replaced by X = do(x) and X = do(x'). This notation helps us to distinguish probabilistic causation studies from standard statistical studies. [8], [12], [17], [16], [18] and [24] have extensive explanations about understanding the concept of interventions and properly representing them in a conditional probabilistic equation. They further go on to suggest some assumptions that need to be made in order to infer the causal effect of one variable on another without conditional bias from other covariates. Below is a brief account of the methodologies proposed in above papers and book to assess interventional effects from observations.

In figure, to denote intervention on variable X to set it to x we have to rewrite the joint distribution as

$$P(A, B, C, D, Y|do(X = x)) \;=\; P(A)P(B)P(C)P(D|C)P(Y|X, D)$$

Every variable connected to the intervened variables is removed from the equation's product. This results from the fact that the post-intervention model (picture the same graph with the causal arrow from A to X removed) is also Markovian. As a result, it must produce a distribution that is factorized using the updated graph to produce the truncated product. All probabilities on the right-hand side are the same as the pre intervention conditional probabilities.

Similarly, the Markov condition factorization (called the truncated factorization) will change to

assuming V = {A, B, C, D, Y};

$$P(A, B, C, D, Y \,|do(X = x)) = \prod_{V} P(V_i \mid Par(V_i))$$

where $P(V_i \mid Par(V_i))$ again, are the conditional probabilities before the intervention.

In a causal relation of the form Z → X → Y, intervening on X to set X = x,

P(Z | do(X=x)) = P(Z) and P(Y | do(X=x)) = P(Y | X=x) where both right-hand side terms are pre-intervention probabilities.

Methodology

The above techniques allow us to study the effect of interventions on outcomes and compare intervention strategies to derive the desired result. Let us follow the step-by-step process to find the causal effects of interventions.

Consider the below graph,

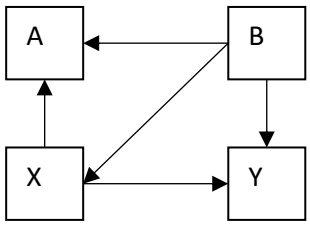


Figure 5: An example causal graph to demonstrate the workings of probabilistic tools of analysing causal relationships.

The causal effect we are interested in is X → Y. Prima facie we can write the joint probability distribution

P(A, B, X, Y) = P(B)P(X|B)P(A|X, B)P(Y|X, B) , and

P(A, B, Y | do(X=x)) = P(B)P(A|X, B)P(Y|X, B) which gives us P(Y | do(X=x)) by marginalizing over A and B

Therefore, $P(Y\ |do(X = x)) = \sum_{A,B} P(B)P(A|X,B)P(Y|X,B)$

A very useful technique to identify for confounders and spurious correlations and selecting variables for adjustment is the backdoor criterion [8] [19] [17]. The causal relation we are interested in is X → Y and is the frontal path. The next step is to trace the paths in the opposite direction from X that end up in Y (called the backdoor). The easiest way to do this is to pick any arrow that enters X and follow along. In the above DAG we have two such backdoor paths.

1. X ← B → Y
2. X → A ← B → Y

These paths need to be blocked as they transfer spurious correlation between the cause and the effect. We see that path 1 contains a confounder and therefore we decide to stratify by the values of B which blocks the path. Path 2 on the other hand contains a collider and therefore blocks causal transmission and hence this path is blocked. We find our sufficient set for adjustment, i.e., stratification by or conditioning upon B that makes sure the relation between X and Y is purely causative.

We can now find causal effects of X on Y by intervention. If we consider a binary state of X = {0, 1}, then the effect of X on Y at a single level or strata of B = b is given by

$$P(Y = 1|do(X = 1), B = b) - P(Y = 1|do(X = 0), B = b)$$

Whereas over the entire population, it is given by a weighted summation over the sufficient set given by

$$\sum_B [P(Y = 1|do(X = 1), B = b) - P(Y = 1|do(X = 0), B = b)]P(B = b)$$

In general, for multivalued variables X and Y, finding a sufficient adjustment set permits us to write

$$P(Y = y|do(X = x), B = b) = P(Y = y|X = x, B = b)$$

And

$$P\big(Y = y\big|do(X = x)\big) = \sum_B P(Y = y, X = x, B = b)P(B = b)$$

Using a suitable set B from the diagram and the back-door criterion, we can immediately write Equation (3.18) without changing the truncated factorization formula. The selection criterion can be used methodically on diagrams of any size and shape, sparing analysts from the challenging mental exercise required by the potential-response framework: determining if "X is conditionally ignorable given B." Additionally, the criterion enables the analyst to look for the best set of covariates for adjustment [16].

To summarize what we know about causation:

1. The idea of causation is born in the mind and strengthened by repeated experiences of events in conjunction.
2. Causality is deterministic in the macroscopic world under ideal conditions, but in a general situation, can also be represented as probabilistic if all the causal factors are unknown. *Causal pluralism.*
3. Causal effects in action over a population can be represented using statistical tools and probability distributions using which, we can deduce information regarding the population.
4. Under non-experimental settings, causal effects can be analysed by prior data observations.
5. It is possible to assess post intervention effects on a causal relation from pre-intervention data observations.

Using our knowledge about probabilistic causation, let's try to analyse a scenario.

Research Objective

Football transfer market

In order to secure the greatest players for the upcoming seasons, the top football clubs in the world compete each year for hundreds of millions of dollars. However, what precisely is a transfer and how does it operate?

Football is a worldwide sport, and elite players frequently change teams. Professional athletes get into agreements with teams for specified terms that might last up to five years. The new club reimburses the old one if a player leaves before the end of their contract. It's called a transfer fee. Twice a year, transfers take place. FIFA rules specify two annual "transfer windows" during which clubs may acquire foreign players. The timing of the two transfer windows varies by country, with the longer one occurring in between seasons and the shorter one in the middle of them. In European footballing nations, the longer transfer window opens in the summer (July–August), whereas the shorter one does so in the winter (December - January). A new contract must be worked out by the athlete, their agency, the club, and all of their attorneys. This includes information on pay and incentives, such as welcome and loyalty bonuses. Players also go through medical evaluations to make sure they are healthy enough to participate. The player does not get the transfer money. Only a small portion goes to the agents who closed the deal. In reality, it's a deal between the player's old club and new club to secure his services. But in essence, this establishes a player's market worth as an asset. There are independent platforms that, using their own algorithms, estimate a player's market value, but actual values are only discovered when there is a transfer that actually occurs that is unrelated to any of these independent projections and clubs will have their own methods of negotiations.

Although the transfer fee does not go to the player, it does make an impression of the player to the world along with setting some expectations of him. A high transfer fee instantly puts the focus of football fans around the world on a player thus increasing his popularity from which, a club may occasionally seek to make a profit. When negotiating contracts, image rights can be a significant subject of contention. The exclusive right to decide how a player's image appears in publicity and advertising (image rights) is typically demanded by clubs. However, players are hesitant to pass up tempting chances to make money through advertisements. Therefore, an agreement must be reached between all parties; for instance, the player and the club may agree to split any profits from the use of the player's image in half.

Nevertheless, a high transfer fee cannot exactly be termed as success for a player, where success is a subjective term in itself. But it will bring the player into the limelight and open up new opportunities which can be termed good for the player’s career. The point we want to check here is that if a player can influence the transfer fee with his performance. The *a priori* notion about the relation between performance and transfer fee is that they are positively correlated. We are attempting to find whether the performance has any causal effect on the transfer fee. If we follow the claim of probabilistic causation, then it should be that if performance causes the transfer fee of a player to be what it is, an increase in performance should increase the probability of a high transfer fee when all back ground conditions are kept fixed as stated in section. The performance that we are talking about can be considered to be an intervention.

To frame our problem in another way, suppose we have a statistical minded football player who wants to secure a high transfer value to his name. But he knows that he can only improve his performance to have any effect on the transfer fee. Yet, he is unsure if that will have any effect because he is aware of several other factors such as his age, the team he plays for (who will get the transfer fee) all might affect the transfer. For example, a younger sportsperson is considered more valuable than one who is aged, or a big team might have upper or lower threshold amount outside of which they will not consider any deal. Among all this, he is apprehensive whether him playing well will have any significant effect on his transfer fee. If it does, then he can say he caused his

transfer fee to be that way. It will also mean that there is evidence for free will. The footballer resorts to probabilistic causation to analyse the causal claim.

Let us follow the process step by step

Step 1: Drawing a DAG to demonstrate causal relationships

A Directed acyclic graph should accommodate all assumed causal relations. The relevance of which can be assessed later. We will start with an extended DAG that attempts to encode within it the causal associations that lead a young footballer to become an older footballer who then makes the transfer to another club.

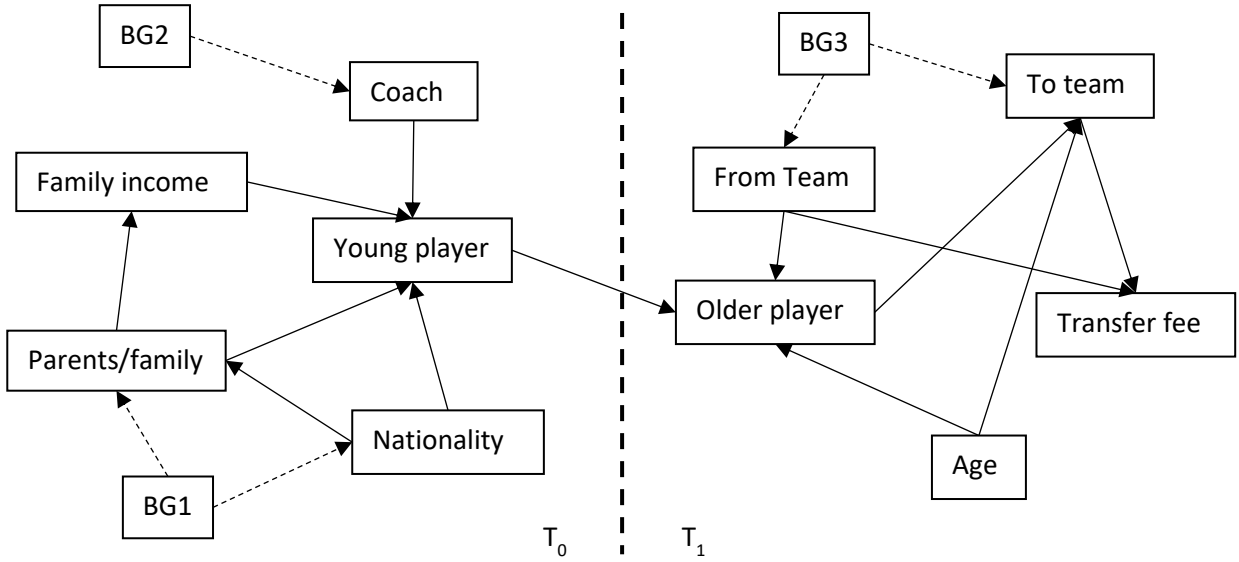


Figure 6: An initial DAG displaying the various factors that may influence the performance and career of a football player

The DAG in figure 6 captures the notion that factors like place of birth (Nationality), family income, family background and coach influence the performance of a young footballer. Whereas factors like the teams involved, age of the player, etc determines the performance of the older footballer. BG1, BG2 and BG3 are unobserved background conditions. T_0 denotes the causal relations of the past while T_1 denotes a future time (whenever the transfer happens). And a player can affect the transfer fee only through *To team.* There is no direct causal effect.

Step 2: Minimizing DAG

The proposed DAG is Markovian as it does not contain any cyclic causal chain. As per Markov condition of screening off, we see that the node *Older footballer* screens off the left side of the DAG from the right side. The idea is that all older footballers are equally likely for transfer and it does not depend on what causal path they took to reach that stage. This means the right side is completely independent of the left side us with a much smaller DAG as in figure

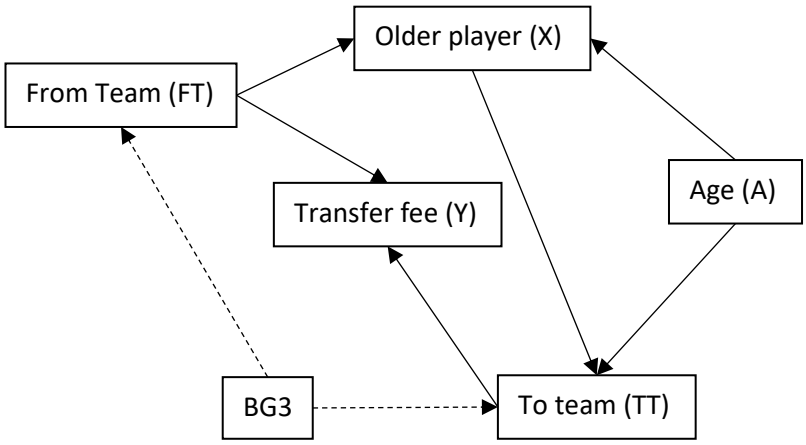


Figure 7: A minimized DAG capturing only the information relevant to the current research

ng

Some explanations are in order regarding the assumptions made. Firstly, age determines a player's fitness and performance and at the same time age (longevity) is something that teams take into consideration while purchasing a player. The old team of the player (FT) causally affects his performance via teammates, club

heritage, coaching staff, etc and may affect transfer fees if they put certain conditions on a transaction. For example, setting a minimum transfer amount. An unobserved background BG3 has also been assumed between FT and TT.

Step 3: Lay out front door and backdoor causal paths to identify confounders and colliders.

The relation we are interested in is X → Y. In this case it is X → TT → Y and it is the only front door path from X to Y. No conditioning is required for this path.

The backdoor paths are as follows:

I. X ← FT → Y
II. X ← FT ← BG3 → TT → Y
III. X ← A → TT → Y

Paths I and II can be blocked by conditioning on FT whereas path III can be blocked by conditioning on A.

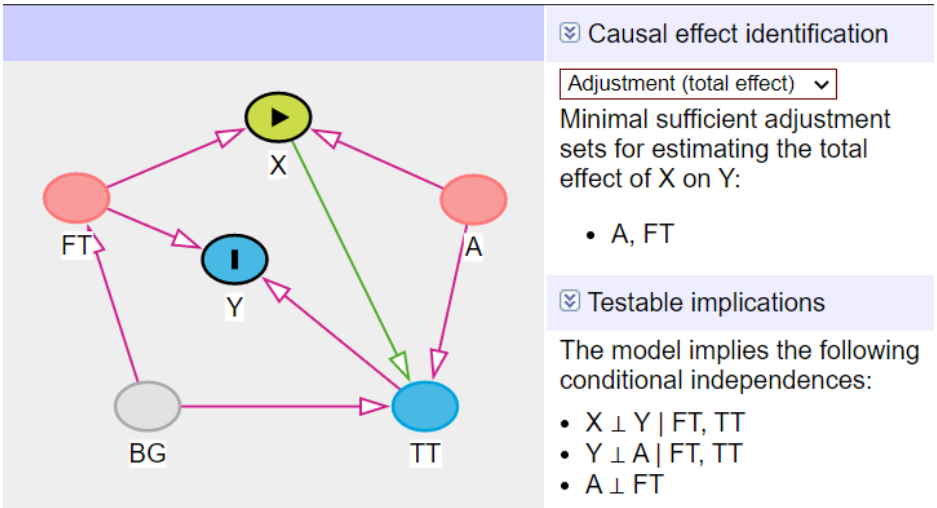


Figure 8: A graphical layout built on dagitty.com. On the right we can see the variables, conditioning for which is a sufficient condition for causality. It matches the observation from path analysis. Dagitty provides some testable implications as well.

Step 4: Perform causal analysis

To make our analysis free from confounding bias, we need to stratify by two variables FT and A, as we saw in the previous step. This method involves holding the confounders fixed at one value which is thus, an intervention. But our data will be observational and not experimental, then how can we obtain the effect of an intervention from an observation. [17] mentions the rules of do-calculus which is of use here. As per do-calculus, actions (interventions) can be replaced by observation under a certain condition (the symbol $\perp\!\!\!\perp$ represents independence).

$$P(y|do(x),do(z),w) = P(y|do(x),z,w) \quad if \quad (Y \perp\!\!\!\perp Z \mid X,W)_{G_{\overline{X}\underline{Z}}}$$

Suppose we club both the confounders in Z such that Z = {A, FT} and the rest of the covariates in W, then the graph $G_{\overline{X}\underline{Z}}$ will look like

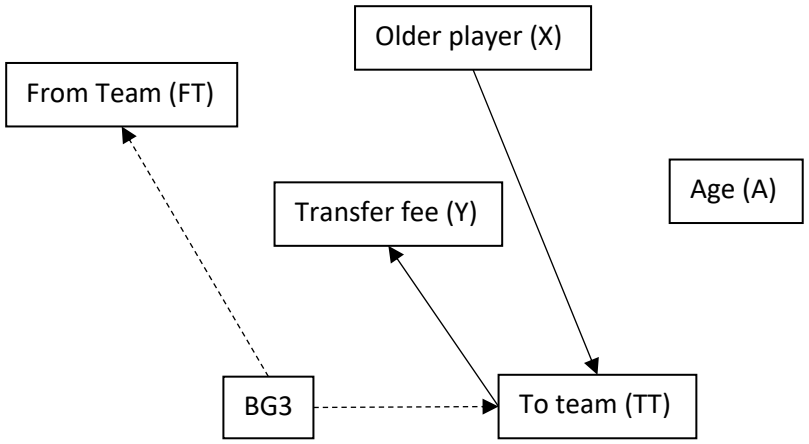


Figure 9: The minimized graph with some arrows removed. Removed arrows demonstrate that the causal connection between variables is broken when one is conditioned upon. Here, variables 'from team' and 'age' are being stratified.

We can see from figure 9 that $Y \underline{||} A$ but $Y \; not \underline{||} \; FT$ and is conditional on unobserved BG3. The work around this, we are considering football transfers only in the top 5 European football leagues, i.e., English, Spanish, Italian, German and French. The assumption is that since these leagues are pretty close knit due to facing off against each other on multiple occasions, they are geographically close and are more popular in the world therefore, the most of the background conditions will be the same for all. This confounding can be therefore considered to be already stratified on a single level. This assumption makes $Y \underline{||} FT$.

Now we can use equations

$$P(Y = y|do(X = x), B = b) = P(Y = y|X = x, B = b)$$

And

$$P\big(Y = y\big|do(X = x)\big) = \sum_{B} P(Y = y, X = x, B = b)P(B = b)$$

And analyse assertions of the essence

$$P(Y = 1|do(X = 1), B = b) - P(Y = 1|do(X = 0), B = b)$$

To find out causal differences between two intervention cases
The workings and notations of the above equations have been already discussed in the methodology section.

Step 5: Collect and process data

This is perhaps the most cumbersome part of the research. Searching for atomic information regarding football on the internet is not hard at all but gathering large amounts of data from different sources on the internet and stitching them together can be quite challenging. After enormous amounts of web scraping, data frame processing in python and some manual adjustments in MS Excel, three master datasets were made.

1. Master player dataset

NAME	POSITION	HEIGHT	FOOT	AGE	RATING	NATION
Mohamed Salah	Attack	175	Left	30	96	Egypt
Cristiano Ronaldo	Attack	187	Right	37	96	Portugal
Kevin de Bruyne	Attack	181	Right	31	96	Belgium
Virgil van Dijk	Defence	193	Both	31	96	Netherlands
Heung-Min Son	Attack	183	Both	30	95	Korea Republic
...	...	...	...	...	...	...
Alaa Bellaarouch	Goalkeeper	191	Right	20	65	Morocco
Mathyas Randriamamy	Goalkeeper	186	Right	19	65	Madagascar
Geoffrey Agbolossou	Goalkeeper	185	Right	22	65	Togo
Yanis Saidani	Goalkeeper	187	Right	20	65	Algeria
Fredler Christophe	Midfield	164	Right	20	63	Haiti

Figure 10: Master player dataset containing all player information

The Master player dataset contains information about all the current players in the top 5 leagues. The data was collected from *en.soccerwiki.org.* This dataset was used to map the player details to their transfer details by their names. The ratings have been calculated by the website's own algorithms which are hidden to users. Although, a different rating scale has been used in the final dataset, obtained from a more comprehensive database. This dataset has been used to estimate the probability distribution of the age variable later on in the analysis.

2. Master transfer dataset

DATE	NAME	FROM	FT_RATING	TO	FEE
19-Jul-22	Ben Davies	Liverpool	6.96	Rangers	Undisclosed
19-Jul-22	Sonny Perkins	West Ham United	6.75	Leeds United	Undisclosed
19-Jul-22	Djed Spence	Middlesbrough	6.59	Tottenham Hotspur	Undisclosed
19-Jul-22	Fábio Silva	Wolverhampton Wanderers	6.74	Anderlecht	Loan
19-Jul-22	Jamie Shackleton	Leeds United	6.65	Millwall	Loan
...	...	...	...	...	...
25-Jan-18	Loïs Diony	Saint-Etienne	6.64	Bristol City	Loan
23-Jan-18	Lassana Diarra	No team	0.00	Paris Saint-Germain	Free
12-Jan-18	Yann M'vila	Rubin Kazan	0.00	Saint-Etienne	Free
04-Jan-18	Prince Oniangué	Wolverhampton Wanderers	6.74	Angers	Loan
03-Jan-18	Yeni N'Gbakoto	Queens Park Rangers	6.62	Guingamp	Undisclosed

Figure 11: Master transfer dataset that contains information about all the transfers in the last 5 years including loans and free transfers.

The master transfer dataset is a compilation of the data collected from www.espn.co.uk and www.transfermarkt.com containing all the player transfers that have taken place in the last 5 years since 2018 till August 2022 among the top 5 leagues. This dataset contains transfers that were free as well as loan transfers. Only complete transfers that cost money have been used in the analysis. FT_RATING that stands for from-team rating has been collected from www.whoscored.com. This dataset has also been used to estimate the probability

distribution of the team ratings of the teams that are left by the players. Only from-team has been considered as it is one of the variables that need to be conditioned upon.

3. Main dataset for analysis

DATE	NAME	RATING	FROM	FT_RATING	TO	TT_RATING	FEE	POSITION	HEIGHT	FOOT	AGE	NATION
19-Jul-22	Ben Davies	7.07	Liverpool	6.96	Rangers	6.80	4.70	Defence	181	Left	26	England
01-Feb-21	Ben Davies	6.69	Preston North End	6.77	Liverpool	6.96	1.85	Defence	181	Left	25	England
13-Jul-22	Raheem Sterling	7.42	Manchester City	7.08	Chelsea	6.89	56.20	Attack	170	Right	27	England
12-Jul-22	Nathan Collins	7.11	Burnley	6.70	Wolverhampton Wanderers	6.74	25.94	Defence	193	Right	21	Ireland
24-Jun-21	Nathan Collins	6.80	Stoke City	6.70	Burnley	6.70	14.65	Defence	193	Right	20	Ireland
...	...	...	...	...	...	...	...	...	...	...	...	...
26-Jun-18	Javier Pastore	7.12	Paris Saint-Germain	7.02	AS Roma	6.81	24.70	Midfield	187	Right	29	Argentina
19-Jun-18	Willem Geubbels	6.28	Lyon	6.87	AS Monaco	6.75	20.00	Attack	185	Right	16	France
19-Jun-18	Issa Diop	6.89	Toulouse	6.60	West Ham United	6.75	25.00	Defence	194	Right	21	France
31-Jan-18	Lucas Moura	6.49	Paris Saint-Germain	7.02	Tottenham Hotspur	6.86	28.40	Attack	172	Right	25	Brazil
25-Jan-18	Guido Carrillo	6.02	AS Monaco	6.75	Southampton	6.66	22.38	Attack	190	Right	27	Argentina

Figure 12: The final dataset formed after combining the previous two and after data pre-processing

This is the main dataset on which the analysis is based. It has been compiled by linking the previous two tables by the player names as the common key. The key steps of pre-processing that were performed are:

- Removal of duplicate rows. Since the transfers are among the top 5 leagues, a single transfer occurred twice in the final dataset, as entries from the two leagues involved in the transfer.
- Updating the ages. Since the ages were mapped from master player dataset, those were the players' current ages. As many years have been subtracted from that number (age) as many years it has been since the transfer happened (date).
- Conversion of currencies. Due to the fact that the data was collected from multiple sources, there were mismatches when it came to units. Dollars, Pounds and Euros have all been converted to million Euros. The exchange rates used were the yearly average exchange rates of the respective currencies in the respective year.
- Handling null values. No data has been imputed in the dataset. Any row missing an important value like rating or transfer fee has been deleted.
- The dataset has not been normalized but the probabilities have been normalized later.

Observations and results

The 5 numerical columns in the main dataset have been used in the analysis. FT_RATING and TT_RATING represent from-team and to-team. The rating adjacent to the team names is the average of the ratings in the last 5 years. The value essentially demonstrates how good a team is in general in its respective league. The player rating (RATING) is the rating in the season before the transfer. This gives an idea about how good a player's performance was right before the transfer. The reason behind this that football clubs are established institutes and one season is not enough to assess its competitiveness on the other hand, we need to know the latest state of the player before the transfer to get an idea on what must have brought it about. All ratings have been collected from www.whoscored.com who also have their own algorithm to rate players. And finally, age and fees.

An initial boxplot shows us the data concentration in a specific field

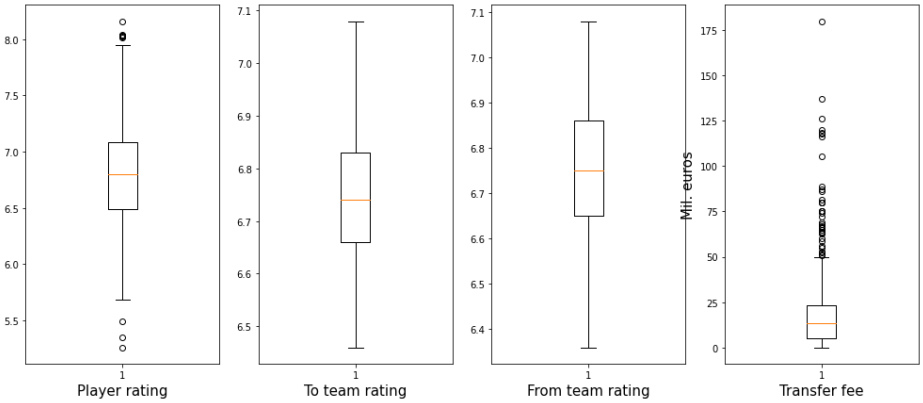


Figure 13: Initial boxplot with outliers

As we can see, there are few outliers in player ratings and several in transfer fee. This tells us that the distribution of the transfer fee is highly skewed. To have a more generalized result, we have to get rid of some of the outliers.

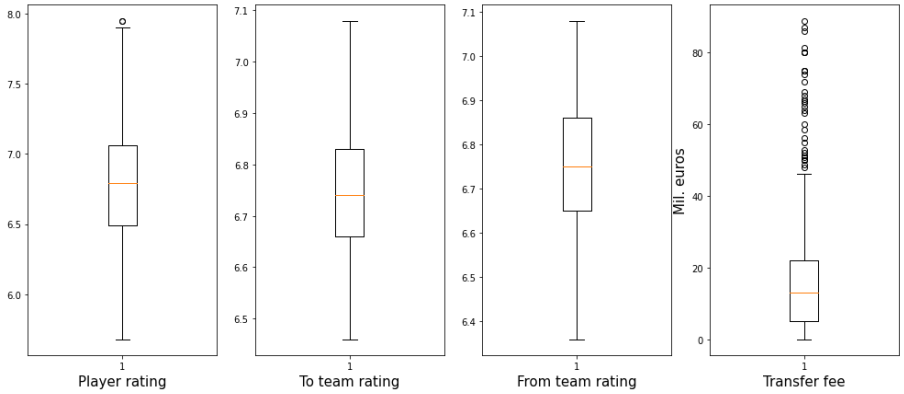


Figure 14: Boxplot with many outliers removed to get generalized results

Now that there are less outliers, we can get a more general result of our analysis. But is the dataset complete for the analysis? Not quite. Since all the numerical variables are continuous (except age), we assign them to four classes (except age) as per the quartile they fall in. We make use of the boxplot's interquartile range to do this. The age classes have been set in 5-year gaps. The class hierarchy works exactly like a ranking system where the best falls under class 1 and worst falls under class 4. This should not cause confusion, for example, for age, the less it is, the better for sports so ages falling under 15 to 20 years are classified as class 1. The opposite is true for the other fields such as rating and fee. Therefore, values falling under the 4^{th} quartile range for these fields have been assigned class 1. Post associating values to their classes the dataframe (figure 15):

DATE	NAME	RATING	P_CLASS	FROM	FT_RATING	T_CLASS	TO	TT_RATING	FEE	TF_CLASS	POSITION	HEIGHT	FOOT	AGE	AGE_CLASS	NATION
19-Jul-22	Ben Davies	7.07	1	Liverpool	6.96	1	Rangers	6.80	4.70	4	Defence	181	Left	26	3	England
01-Feb-21	Ben Davies	6.69	3	Preston North End	6.77	2	Liverpool	6.96	1.85	4	Defence	181	Left	25	2	England
13-Jul-22	Raheem Sterling	7.42	1	Manchester City	7.08	1	Chelsea	6.89	56.20	1	Attack	170	Right	27	3	England
12-Jul-22	Nathan Collins	7.11	1	Burnley	6.70	3	Wolverhampton Wanderers	6.74	25.94	1	Defence	193	Right	21	2	Ireland
24-Jun-21	Nathan Collins	6.80	2	Stoke City	6.70	3	Burnley	6.70	14.65	2	Defence	193	Right	20	1	Ireland
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
26-Jun-18	Javier Pastore	7.12	1	Paris Saint-Germain	7.02	1	AS Roma	6.81	24.70	1	Midfield	187	Right	29	3	Argentina
19-Jun-18	Willem Geubbels	6.28	4	Lyon	6.87	1	AS Monaco	6.75	20.00	2	Attack	185	Right	16	1	France
19-Jun-18	Issa Diop	6.89	2	Toulouse	6.60	4	West Ham United	6.75	25.00	1	Defence	194	Right	21	2	France
31-Jan-18	Lucas Moura	6.49	4	Paris Saint-Germain	7.02	1	Tottenham Hotspur	6.86	28.40	1	Attack	172	Right	25	2	Brazil
25-Jan-18	Guido Carrillo	6.02	4	AS Monaco	6.75	3	Southampton	6.66	22.38	1	Attack	190	Right	27	3	Argentina

Figure 15: Final dataset after continuous numerical values have been associated with a given class

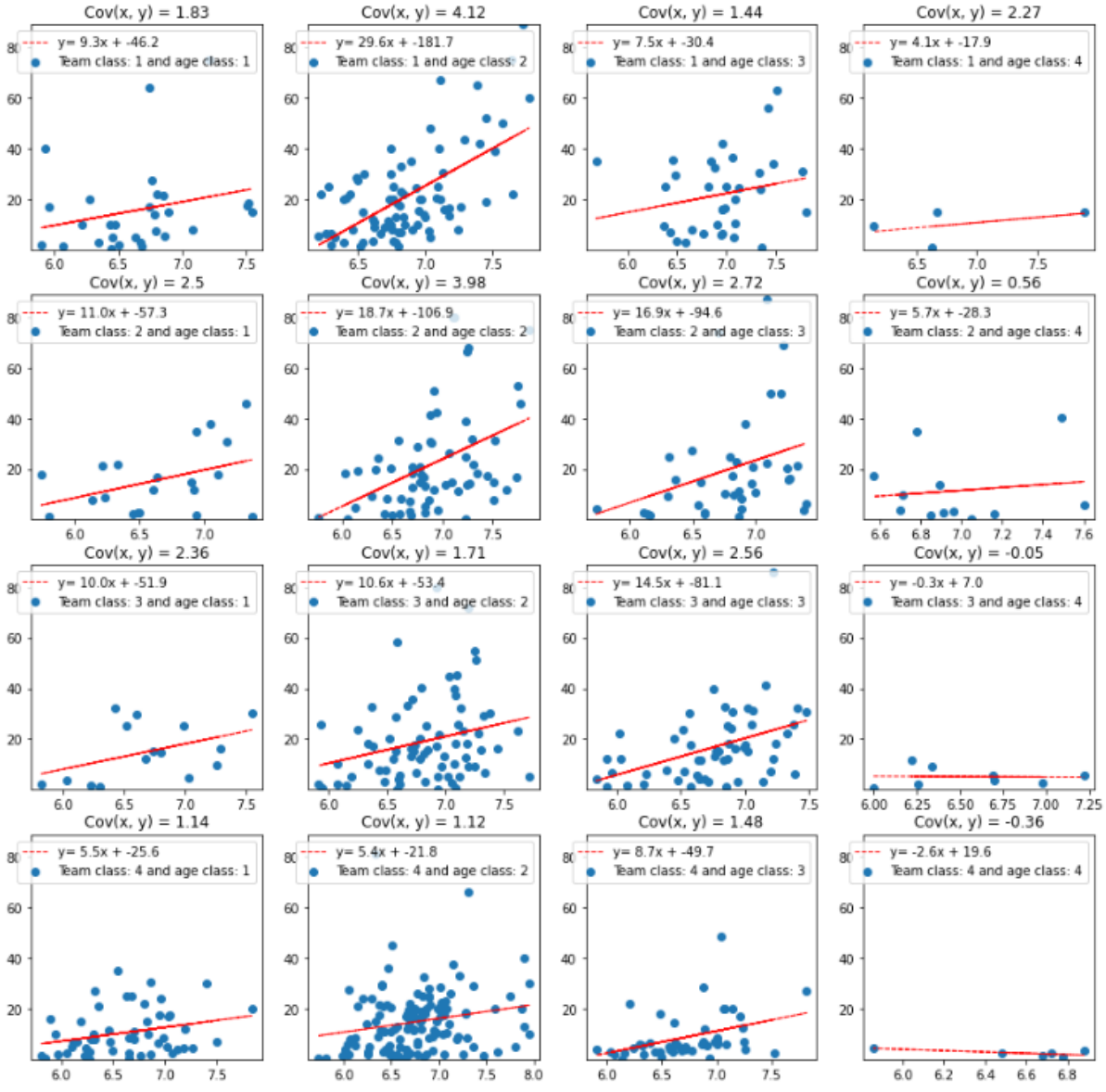


Figure 16: A 4x4 grid display showing the correlation between player rating (performance) and transfer fee for every pair of age and team class

In figure 16, we can see the transfer fee plotted against player rating at every level of stratification of the confounding variable. There are some observations to be made here. First, age class 2 and 3 are represented more than the other classes. This gives us an idea that players within that age range are more in the profession of football and that transfers mostly happen at that age. This does not imply anything regarding causation but puts a constraint to the question of intervention we will answer later in the paper. Another interesting thing to note is that transfers scatter plots from team classes 2 and 3 are less populated than those from team class 1

and 4. These two observations do not necessarily add anything to our current research apart from emphasizing the effects of confounding and the need to condition on them. In order to visualize the causal influences better, trend lines and covariance between x and y have been provided in the legend.

Our task involves estimating the probabilities of each class for the fields AGE_CLASS and T_CLASS. We have done this using the continuous data in their parent fields AGE and FT_RATING respectively by using Bayesian inference.

We have fit the values in AGE to a gamma distribution with a half normal distribution as prior with a large standard deviation of 10 for both alpha and beta. This is because of the prior knowledge that both will be positive. Bayesian inference in python has been performed using the pymc3 library. The results are shown in figure 17.

The corresponding probability distribution has been plotted. Multiple plots with parameters randomly selected

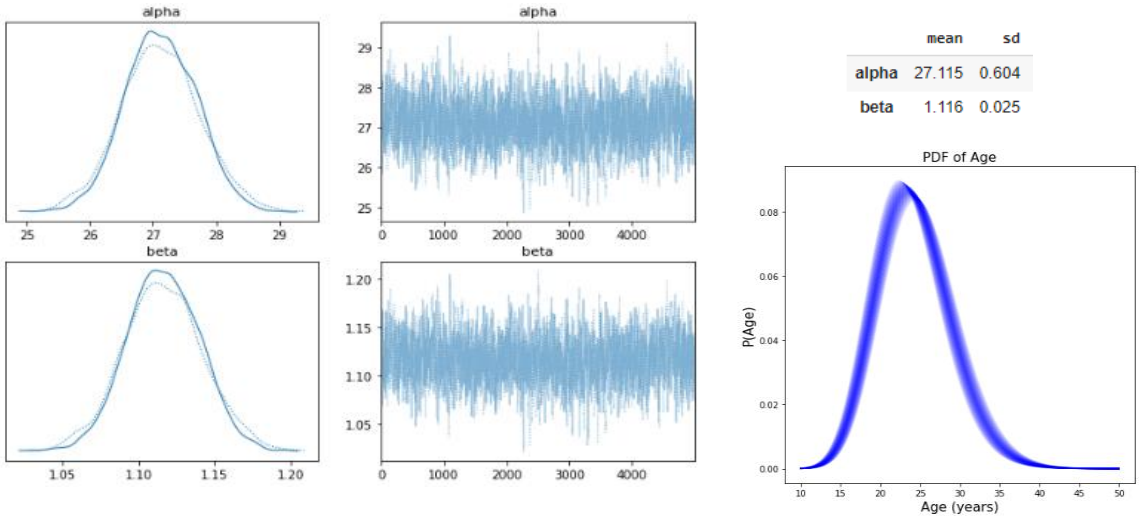


Figure 17: A set of graph plots showing the estimation of the pdf of the age variable. On the left is the output from pymc3 showing estimated value of the parameters. On top right are same the values. The bottom right is a plot of the pdf using values from pymc3

within the resulting standard deviations have been superimposed to visualize the error margins. Bayesian inference gives us a tolerant result with some margin of error in the distribution parameters. Whereas other methods such as distfit give us the best fitting parameters. For example, distfit gives the result in figure 18.

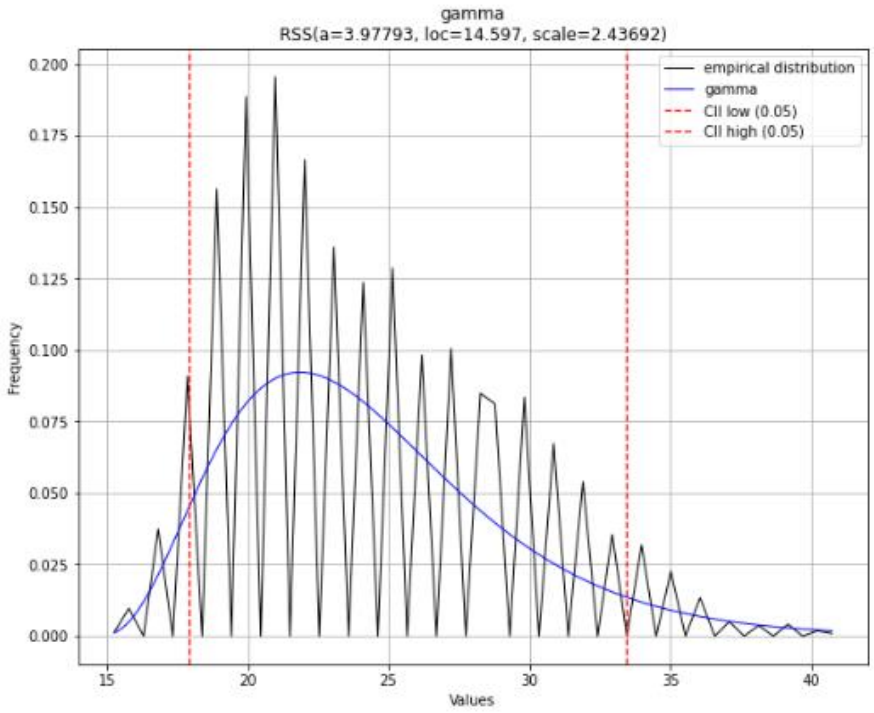


Figure 18: Distfit library's result of fitting ages to a gamma distribution

We have found the probabilities of the age classes by integrating the probability distribution function within the ranges that they are defined and normalizing them. The results are as follows:

```
P(Age class 1) = 0.18079436412608826
P(Age class 2) = 0.40309575054680785
P(Age class 3) = 0.2999378343672786
P(Age class 4) = 0.1161720509598342
Sum of probabilities = 1.0000000000000089
```

The same steps have been followed for estimating the pdf of FT_RATING. A slight difference is that the fit is for a beta distribution so first the values have been scaled down to between 0 and 1. The remaining process is identical. The results of which are shown in figure 19:

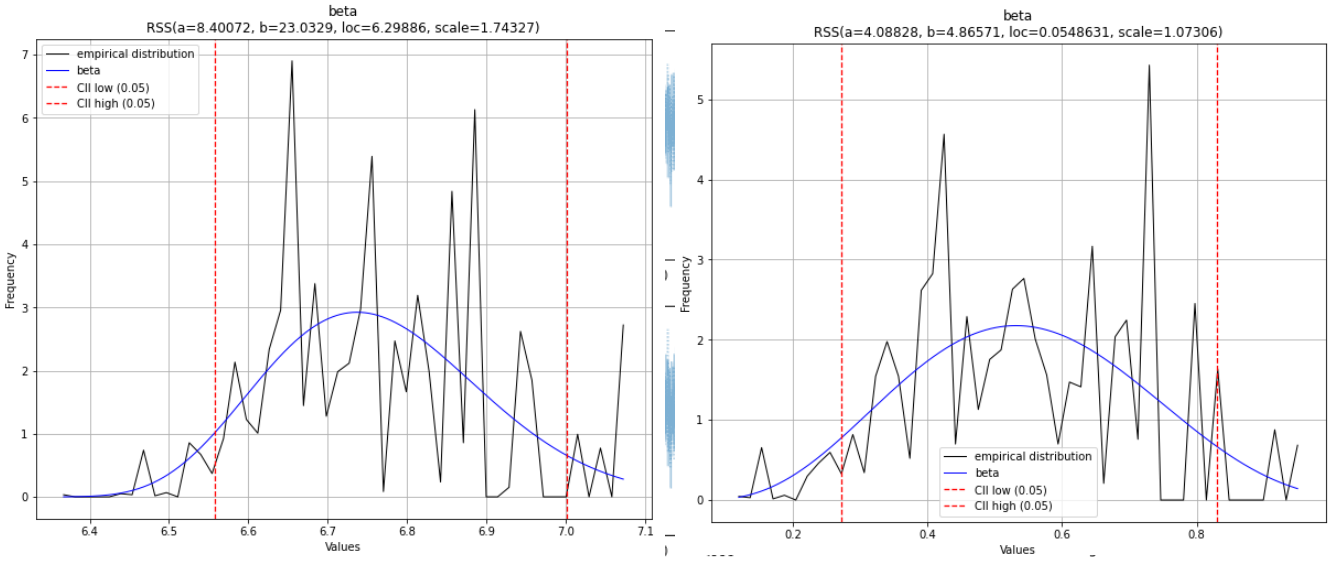


Figure 19: Pymc3 evaluation of parameters for the distribution of from team ratings. On right bottom is the superimposed plot of multiple beta distributions with parameters randomly chosen within the resulting standard deviation (top right)
Figure 20: Distfit plot of team rating fitting to a beta function. The left plot fits the original values whereas the right one fits values scaled down to between 0 and 1

And similarly, the probabilities for from-team class are as follows:

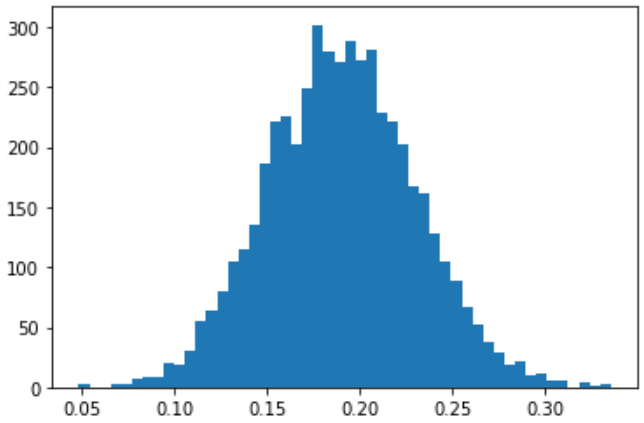
```
P(Team class 1) = 0.2177254861940799
P(Team class 2) = 0.30456322711265943
P(Team class 3) = 0.26196763034264775
P(Team class 4) = 0.21574365634960754
Sum of probabilities = 0.9999999999989946
```

Using these values, we now apply $P(Y = y|do(X = x)) = \sum_B P(Y = y, X = x, B = b)P(B = b)$ which gives us

	Fee class 1	Fee class 2	Fee class 3	Fee class 4
Player class 1	0.38860	0.27308	0.20051	0.11271
Player class 2	0.25888	0.27640	0.23091	0.20848
Player class 3	0.15470	0.22001	0.28741	0.33787
Player class 4	0.15492	0.16027	0.25789	0.39149

Figure 21: The total causative effect over all levels of stratification from each player class to each transfer fee class. The values show strong correlation between performance and transfer fee

A correlation can be seen between a high-class player securing a high-class transfer fee and vice versa. Now that we have found the causal influence of individual interventions, we can now answer the relative causal influence of two interventions. For example, P(Y=1|do(X=1)) – P(Y=1|do(X=2)) ≈ 0.13 which means that a class 1 player has 13% more chance of securing a class 1 transfer fee than a player from class 2. But is this difference of any significance or did it come by chance? To test this, we calculate the difference P(Y=1|do(X=1)) – P(Y=1|do(X=not 1)) on random 400 samples from our main dataset over all stratification levels. We take 5000 such differences and perform z test. We consider a null hypothesis that the difference is 0.0 and any observable difference is only by chance. Using statsmodels library in python, we get a p value of 0.0 and therefore we reject the null hypothesis. This means that there is definitely causal effect of player ratings and player performance on the transfer fee.




```
[69] ztest_Score, p_value= ztest(diff,value = 0.0)

[70] p_value

     0.0

[71] ztest_Score

     337.62769523641515
```

Figure 22: the sampling distribution of the difference in causal influence when treatment/input do(x=1) and do(x= not 1). On the right, it shows the significance of the difference. A p-value of 0.0 allows us to reject the null hypothesis that there is no causal difference between do(x=1) and do(x= not 1).

On average, it was found that, P(Y=1|do(X=1)) – P(Y=1|do(X= not 1)) ≈ 0.19 which is even more than the previous result. This means that a player of class 1 has the highest chance of securing a class 1 transfer fee. So, if somehow, the player can attain the performance of a class 1 player, it maximizes the chance for him to secure a high transfer fee. If the player is improving his performance, can it be called an intervention? It can definitely be called free will.

We then plotted the above difference for each class of the confounders and plotted their values in figure 23.

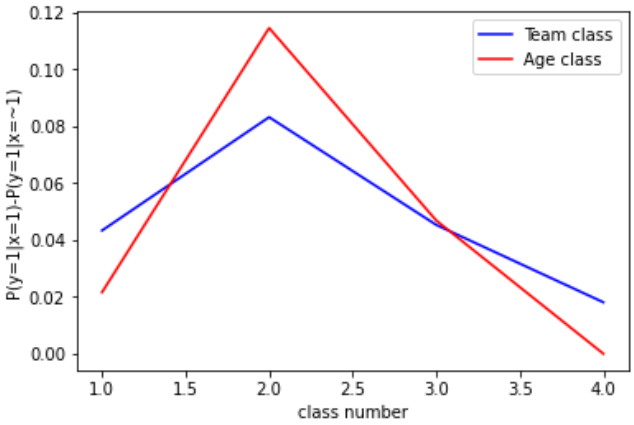


Figure 23: Shows the causal influence difference upon an action do(x=1) and do(x=~1) for every class value of the confounding variables age and from-team

This plot shows the influence of the class 2, towering over everything else. This could mean that the difference of P(Y=1|do(X=1)) – P(Y=1|do(X= not 1)) might only be relevant under certain conditions meaning that there must be bounds to free will or even probabilities associated with it too.

Discussion

To summarise the work, we have presented the reader with an account of the historical and traditional notion about causality, emphasizing on its main beliefs and the scepticism surrounding probability theories. We then pointed out scenarios in which the traditional view on causality finds it difficult to accommodate for uncertainties revolving unobserved background conditions and intrinsic randomness in a system. We then presented the reader with the probabilistic notion that causes raise the probability of an effect. To which we have cited other works and provided a full-length discussion on the place of probability in causation. We have referred to Judea Pearl's Bayesian network or graphical model or DAGs to capture the workings of causation that makes it easier for the reader to grasp the concept. All mathematical formulae require a knowledge of conditional probabilities but are not overly complicated. For more intrinsic knowledge, one can read further into cited works. In our research objective, we presented a scenario where we had the opportunity to utilize the concepts learnt in the literature review. In which, we analysed football transfers and if a player's performance had any causal influence on the transfer fee. The results show that they do. Based on which we study if a player's own intervention in his performance, i.e., the conscious act of *doing* has any causal effect on his transfer value. If it does, then it could be considered as evidence of free will. The results showed that the difference between causal influences of different interventions to the same outcome is not just by chance, which means the differences are significant. But under certain conditions only.

Conclusion

The deep-rooted notion of causality still seems to elude our understanding. The claim that events cannot explain causation but events *can be* explained in terms of causation seems to be true. But the ability to represent natural occurrences, the normal flow of events (and occasional interventions) in a graphical and statistical framework is a marvellous feat and everyone who has contributed to that work deserve praise. This ability allows us to study occurrences in the world that are invisible to the eye and only the data can tell. This ability helps us make even better inferences or predictions about something the next time. It can be, quite debatably, said that this goes back to Humean causality that claims humans learn to associate events only by repeated observations, only now our observations aren't limited to the bodily senses. Our tools to identify a real cause from a false has improved considerably, due to which we are able to implement causal notions about the functioning world into AI algorithms and let them learn to identify the relevant factors from the irrelevant. The ability of statistics to emphasize unseen, even sometimes incomprehensible, relations is perhaps its biggest strength and arguably the probabilistic notion of causation, of inferring from samples may not be mathematically rigorous and be subject of scepticism, but it is one that is embedded in our consciousness. As for free will, it does exist but only in closely knit and localized scenarios like in the social framework but otherwise, the universe seems to be caused by higher deterministic and probabilistic forces.

Bibliography

[1] D. Hume, A Treatise of Human Nature, 1740.

[2] C. M. Lorkowski, “David Hume: Causation,” *Internet Encyclopedia of philosophy,* 2011.

[3] K. A. Rogers, “Hume on Necessary Causal Connections,” *Philosophy,* 1991.

[4] H. a. M. G. Andreas, “Regularity and Inferential Theories of Causation,” *The Stanford Encyclopedia of Philosophy,* 2021.

[5] C. McLear, “Kant: Philosophy of Mind,” *Internet Encyclopedia of Philosophy,* 2015.

[6] J. E. Moyal, “Causality, Determinism and Probability,” *Cambridge University Press on behalf of Royal Institute of Philosophy,* 1949.

[7] T. I. Philosopher, “Causality,” [Online]. Available: https://www.informationphilosopher.com/freedom/causality.html.

[8] C. Hitchcock, “Probabilistic Causation,” *The Stanford Encyclopedia of Philosophy,* 2021.

[9] I. Hacking, The Emergence of Probability: A Philosophical Study of Early Ideas about Probability, Induction and Statistical Inference, Cambridge University Press, 2006.

[10] T. M. Porter, “Probability and statistics,” *Encyclopedia Britannica,* 2020.

[11] J. Bernoulli, Ars Conjectandi, 1713.

[12] P. J.-L. Eugenia Goldvarg, “Naive causality: a mental model theory of causal meaning,” *Cognitive Science,* 2001.

[13] P. J.-L. Caren A. Frosch, “Is everyday causation deterministic or probabilistic?,” *Acta Psychologica,* 2011.

[14] J. S. Pengfei Yin, “Is causation deterministic or probabilistic? A critique of Frosch and Johnson-Laird,” *Journal of Cognitive Psychology,* 2011.

[15] J. Pearl, Probabilistic reasoning in Intelligent systems, 1986.

[16] J. Pearl, “Statistics and Causal Inference: A Review,” *Sociedad de Estad´ıstica e Investigaci´on Operativa,* 2003.

[17] J. Pearl, “A Probabilistic Calculus of Actions,” 2013.

[18] R. H. Benjamin Margolin Rottman, “Reasoning About Causal Relationships: Inferences on Causal Networks,” *American Psychological Association,* 2013.

[19] A. Heiss, “Andrew Heiss,” [Online]. Available: https://www.andrewheiss.com/blog/2021/09/07/do-calculus-backdoors/. [Accessed 25 08 2022].

[20] J. P. Jin Tian, “Probabilities of causation: Bounds and identification,” *Annals of Mathematics and Artificial Intelligence,* 2000.

[21] H. Reichenbach, The Direction of Time, Dover Publications, 1956.

[22] N. Cartwright, “Causal Laws and Effective Strategies,” *Noûs,* 1979.

[23] B. Skyrms, “Causal Necessity,” *Philosophy of Science,* 1981.

[24] R. McElreath, Statistical Rethinking, CRC Press, 2015.