Trinity College Dublin
Coláiste na Tríonóide, Baile Átha Cliath
The University of Dublin

PhD Thesis

Trinity College Dublin

The University of Dublin

Business Analytics

The journey towards discovering a novel marketing attribution method.

Author:

Georgios Filippou

21367214

Supervisor:

Dr. Ashish Kumar Jha

Submitted

July 17, 2026

A dissertation submitted in partial fulfillment of the degree of Doctor in Philosophy

Declaration of Work

I hereby declare that this dissertation is entirely my own work and that it has not been submitted as an exercise for a degree at this or any other university.

I have read and I understand the plagiarism provisions in the General Regulations of the University Calendar for the current year, found at http://www.tcd.ie/calendar.

I have completed the Online Tutorial on avoiding plagiarism 'Ready Steady Write', located at http://tcd-ie.libguides.com/plagiarism/ready-steady-write.

I consent to the examiner retaining a copy of the thesis beyond the examining period, should they so wish (EU GDPR May 2018).

I agree that this thesis will be publicly available, subject to Irish Copyright Legislation and Trinity College Library conditions of use and acknowledgement.

Signed: ____________________ Date: ____________________

Abstract

The global advertising industry is projected to reach \$2.5 trillion by 2030. It reflects the growing dependence of businesses on media and advertising platforms to attract and retain customers. The rise of advertising as a dominant business model, particularly among large digital platforms, has significantly increased the cost of market access. As a result, businesses left with little choice but to invest across a growing number of marketing channels. This challenge is particularly prominent for digital-native businesses seeking broad audiences in competitive markets. Despite advances in research and technology, smaller-scale advertisers remain poorly equipped to answer one of the most fundamental questions in marketing: what is the best use of our advertising budget? This thesis addresses this gap through three empirical studies. The first investigates whether web traffic generated by various marketing channels influences direct traffic and purchase behaviour, establishing the interdependencies that exist across channels. The second examines whether digital media investments generate spillover effects on organic branded searches, revealing how paid activity shapes unpaid performance. The third, driven by the findings of the first two, proposes a causal-driven attribution framework that estimates channel influence without relying on user-level data, offering a privacy-compliant and accessible measurement approach for businesses of all sizes. Together, these studies contribute to a more rigorous and equitable understanding of multi-channel marketing measurement. This work contributes to both marketing practice and academic literature. For practitioners, it offers actionable frameworks to measure channel performance, allocate budgets more effectively, and make evidence-based decisions without requiring technical expertise or user-level data. For researchers, it advances the understanding of multi-channel attribution, media spillover effects, and causal inference in advertising measurement, areas where empirical evidence remains limited.

Acknowledgements

And here we are. On a Saturday afternoon, tackling the most emotional task of this years-long process: writing the acknowledgements. I believe this is a moment of reflection. Regardless, I am treating it as such.

The present work describes a journey. A journey that, through three published empirical studies and three unpublished ones, provides evidence of my academic experience. To the future reader: it was not a pleasant journey. But as scientists, researchers, professionals, and individuals, we do what we have to do.

I have always considered myself a marketing practitioner. Through work and studies, I had the opportunity to immerse myself in the wonderful world of statistics, probabilities, and causal inference. Naturally, I also engaged with time-series data since marketing is all about it.

Many years ago, while working full-time, I stumbled upon a problem that all marketing managers face: a large percentage of sales was coming through direct traffic. Users were purchasing during their first visit to a website, arriving directly. No one around me seemed to be bothered by this. But it stuck with me. It bothered me. I felt there was something bigger out there to discover.

Almost eight years later, I realise that was the beginning of my academic journey.

Then came the applications, the paperwork, the interviews, all to get into a part-time PhD programme in the UK. Then, unexpectedly, I relocated to Ireland, and had to pursue my studies while working full-time.

The first draft of the first paper took shape while the world was learning what a pandemic was. And as every journey has its ups and downs, I had to find a new university, as the pandemic and Brexit made it impossible for EU students to pursue a part-time PhD in the UK while living abroad.

Then came Trinity, following a six-month search for a new supervisor.

The first publication was a success. I finally managed to demonstrate that users exposed to marketing messages do indeed end up visiting business websites directly. We uncovered an indirect link between advertising channels and the largest source of traffic: Direct.

The second idea followed naturally. If channels influence one another, how do paid advertising channels impact branding channels within the digital ecosystem? After a few rejections, we published our work and established an indirect effect of paid ads on branding objectives.

But the journey was not over. As with every good story, success was followed by challenges. I had to find a new supervisor. Luckily enough I did.

While trying to launch a new venture, the third publication was taking shape. Having established how users behave and shop, I needed a framework that could quantify this. Not only for myself and my new venture, but for everyone. And so the CDA (Causal-Driven Attribution) was born. A novel approach that went above and beyond existing methods: a data-driven attribution model that has received considerable attention and is changing the way the industry treats data.

I feel accomplished.

To my wife Claire: You believed in this more than I did on the days it mattered most. This is as much yours as it is mine. None of this exists without you

To my side of family: Thank you. You supported me more than you know.

To my wife's family: You welcomed me, supported me, and never once made me feel like this journey was mine alone. I am more grateful than words allow.

To my friends: You never fully understood what this journey involved, and yet you were always there. That is the best kind of support.

To my team and industry engagements: This work lives at the intersection of practice and theory, and that is largely because of you. Your influence is embedded in every page.

To my supervisor Professor Ashish: Few words can capture what your support has meant. You kept this journey alive when it could have easily ended. I will always be thankful to your guidance and leadership.

Contents

1 Introduction **1**

1.1 Background and motivation 1

1.2 Aims and objectives 4

1.3 Contributions 9

1.3.1 Empirical Contributions 10

1.3.2 Methodological Contributions 11

1.3.3 Theoretical Contributions 12

1.3.4 Practical Contributions 13

1.4 Structure of the Thesis 14

2 Establishing the link: Does web traffic from various marketing channels influence direct traffic source purchases? **3**

2.1 Introduction 3

2.2 Theoretical background 5

2.2.1 Marketing synergies and cross-media 5

2.2.2 Web attribution 5

2.2.3 Consumer journey 6

2.3 Hypothesis development 7

2.4 Methods 8

2.4.1 Data 8

2.4.2 Model 9

2.5 Results 9

2.5.1 Model evaluation 11
2.6 Discussion and conclusion 13

3 Unveiling digital dynamics: Do digital media investments impact organic branded searches? 15

3.1 Introduction 16
3.2 Theoretical background 20
3.2.1 Consumer journey 20
3.2.2 Attribution modelling 22
3.2.3 Marketing synergies 23
3.3 Data 25
3.4 Methodology 26
3.4.1 Ridge regression 26
3.4.2 Adstock transformation 27
3.4.3 Model estimation 28
3.5 Results 29
3.6 Discussion and conclusion 32
3.6.1 Managerial implications 34
3.6.2 Research contributions 35
3.6.3 Limitations and future directions 35

4 Causal-driven attribution (CDA): Estimating channel influence without user-level data. 37

4.1 Introduction 38
4.2 Literature Review 41
4.2.1 Marketing Effectiveness and the Consumer Journey . 41
4.2.2 Marketing Attribution 42
4.2.3 The Causal Imperative 45
4.2.4 Toward a New Paradigm 46
4.3 Methodology 47
4.3.1 Data Generation 47
4.3.2 Modelling 53
4.3.3 Evaluation Metrics 59

4.4 Results 66
4.4.1 Optimal Time-delay Causality 66
4.4.2 Causal-Driven Attribution Analysis 68
4.5 Discussion and Conclusion 74
4.5.1 Theoretical Implications 77
4.5.2 Managerial Implications 79
4.5.3 Limitations and Directions for Future Research . . . 80

5 Conclusion **83**
5.1 Overview 83
5.2 The Journey: From Pattern Discovery to Causal Attribution 84
5.2.1 Step 1: Finding the Hidden Influence of Marketing Channels on Direct Traffic 84
5.2.2 Step 2: Uncovering the Brand Effects of Performance Campaigns 85
5.2.3 Step 3: From Correlation to Causal Attribution . . . 85
5.3 Summary of Methods and Data 87
5.4 Collective Contribution 87
5.5 Future Research Directions 88
5.5.1 Improving Causal Discovery Accuracy 88
5.5.2 Identifying Other Common Patterns of Consumer Behaviour 89
5.5.3 Making Attribution Decisions Without User-Level Data 89
5.5.4 Scaling to Real-World Validation 90
5.6 Closing Remarks 90

Bibliography **116**

List of Figures

1.1 Research roadmap illustrating the four sequential objectives of the thesis, from establishing cross-channel interdependencies to validating a privacy-compliant causal attribution framework. 2

2.1 Dependent and independent variables (scatter plots) 10
2.2 Fitted vs. Residuals and Normal Q–Q plot, Model 10 12
2.3 Fitted vs. Residuals and Normal Q–Q plot, Model 10 without outliers 12

3.1 Fitted vs. Residuals plot, baseline linear regression model . . 30
3.2 Normal Q–Q plot, baseline linear regression model 30
3.3 Ridge trace plot, $\lambda = 191.048$ 31
3.4 Fitted vs. Residuals plot, final ridge regression model with adstock 32

4.1 Example of a DAG's structure, including nodes: *Affiliates*, *Facebook*, *TikTok*, *Google Ads*, *YouTube*, and *conversion*. Edges represent directed causal influences: green arrows indicate positive values (putative amplification/halo), and red arrows indicate negative values (putative substitution/crowd-out). Nodes are grouped into causal layers as follows: S^3: *Affiliates*, S^2: *Facebook*, *TikTok*, *Google Ads*, S^1: *YouTube*, S^0: *conversion*. 49

4.2 Example of a causal DAG showing the structural relationships among marketing channels. Some channels act as upstream drivers that influence conversions indirectly through intermediate touchpoints, while others have a direct impact on the outcome. All effects flow in a single direction with no circular feedback, providing a clear view of how influence moves through the system. 50

4.3 Overview of the simulation framework. 53

4.4 A random simulated data illustrating marketing channel impressions and conversions within 365 days. The top panel presents the impression trends for five marketing channels, including Google Ads, Affiliates, TikTok, Facebook, and YouTube. Each series follows a distinct stochastic pattern generated from a linear growth process with random noise and inter-channel influence. 54

4.5 Example DAG output illustrating multi-channel interactions from simulated data. Nodes represent marketing channels (Affiliates, Google Ads, TikTok, Facebook, YouTube) and the target outcome (conversions). Directed arrows denote causal influence from cause to effect, with green and red arrows indicating positive and negative causal weights, respectively. 54

4.6 Kernel density estimates (KDEs) of six evaluation metrics, including AUC, TPR, FPR, F0.5, SHD, and SID across lag values (τ). Each subplot shows how the empirical distribution of the metric varies as τ increases from 0 to 60, using 1000 simulation seeds for each τ. The viridis color gradient is used to visually distinguish KDE ridges at different lag values; color does not encode density or any statistical quantity. Instead, the horizontal width of each ridge reflects the density and spread of metric values at a given lag. The red dashed vertical line marks $\tau = 45$, and the red dashed horizontal line denotes the corresponding median metric value at that lag. 67
4.7 Comparison between the ground-truth DAG (a) and the DAG predicted using PCMCI (b). 69
4.8 Comparison of causal effect estimates and residuals for (a) the predicted causal graph and (b) the simulated ground-truth causal graph, compared to the true causal effects. Each panel shows the estimated causal effects per marketing channel and the residual differences between estimated and true values. . 71

1 Introduction

1.1 Background and motivation

Marketing activities are essential for businesses (Urbonavičius and Dikčius, 2008). At its core, marketing enables organisations to communicate value to existing and prospective customers, differentiate themselves in competitive markets, and sustain long-term growth (Kotler et al., 2016). The strategic importance of this function is reflected in the scale of investment it commands globally: advertising spend is projected to exceed $1.25 trillion in 2025 (Statista, 2025*a*). Research demonstrates that a strong and influential marketing department contributes positively to firm performance, regardless of firm size (Wirtz et al., 2014).

Despite the scale of these investments, the relationship between marketing expenditure and business outcomes remains notoriously difficult to quantify. The fundamental challenge, long established in the literature, is that consumers do not respond to marketing stimuli in isolation. They encounter marketing messages across search engines, social platforms, video environments, email campaigns, and offline channels before arriving at a purchase decision (Schweidel, Bart, Inman, Stephen, Libai, Andrews et al., 2022; Verhoef et al., 2015). This fragmentation means that the effect of any single marketing activity is entangled with the effects of all others, making straightforward measurement effectively impossible without appropriate methods and tools (Naik and Raman, 2003; Naik and Peters, 2009). The difficulty is compounded by temporal dynamics: advertising effects do not occur instantaneously but carry over across days and weeks, gradually decaying in their influence on consumer behaviour (Jin et al., 2017; Ephron and McDonald, 2002).

These measurement challenges are not merely academic. Organisations that cannot reliably attribute outcomes to specific activities risk systematically misallocating their budgets, overinvesting in channels that appear productive only because they are easy to measure, and underinvesting in channels whose contributions are indirect or delayed (Danaher and van Heerde, 2018). Attribution modelling defined as the practice of assigning credit for observed outcomes to the marketing touchpoints that preceded them. It has therefore emerged as a central tool for marketing accountability (Buhalis and Volchek, 2021; Kannan et al., 2016). Yet despite decades of methodological development, from heuristic rule-based models to data-driven multi-touch and causal approaches, attribution remains what Ben Mrad and Hnich (2024) describe as one of the most fundamental problems facing modern marketers.

The growing complexity of the digital landscape has made this problem considerably harder to solve. Consumers now navigate omnichannel environments in which the boundaries between paid, earned, and owned media are increasingly blurred (Lemon and Verhoef, 2016). Standard click-based attribution system is the backbone of most commercial analytics platforms. It captures only a fraction of this complexity, recording the last measurable click before conversion while ignoring the upstream exposures and cross-channel interactions that may have been decisive in shaping purchase intent (Dalessandro et al., 2012; Romero Leguina et al., 2020). Furthermore, evidence suggests that impressions marketing exposures that do not generate a click can materially influence consumer behaviour. That includes driving branded keyword searches and ultimately facilitating conversions (Ghose and Todri-Adamopoulos, 2016). These impression-driven pathways are structurally invisible to click-centric attribution frameworks, introducing systematic bias into measurements of channel effectiveness.

The situation has been further complicated by a sharp contraction in the availability of user-level data. Privacy regulations including the European Union's General Data Protection Regulation (Li et al., 2019*a*), the California Consumer Privacy Act (Bonta, 2022*a*), and platform-level changes such as Apple's iOS 14 updates (Sokol and Zhu, 2021) have significantly restricted the cross-platform tracking upon which most data-driven attribution models depend. The deprecation of third-party cookies has reinforced this trend (Oksanen, 2022; Kamena, 2021). Collectively, these developments have created what might be termed a privacy-attribution paradox: the analytical need for sophisticated multi-touch attribution is growing precisely as the data infrastructure required to support it is becoming less accessible (Wernerfelt et al., 2025).

These challenges are felt unevenly across the business landscape. Large corporations with dedicated data science teams, substantial technical infrastructure, and long-established measurement frameworks are better positioned to adapt. Small and medium-sized enterprises (SMEs), by contrast, operate under significant resource constraints. Yet SMEs constitute the overwhelming majority of businesses in most economies. They account for over 99% of all enterprises in the European Union and representing approximately two-thirds of total private sector employment (European Commission, 2023). They are active participants in digital advertising markets, allocating meaningful portions of their limited budgets across social media, search engine marketing, display networks, and programmatic channels in pursuit of growth (Leeflang et al., 2014). Indeed, SMEs are frequently early adopters of emerging marketing channels, drawn by lower entry costs and direct access to audiences that would otherwise be beyond their reach (Qin et al., 2024).

For these businesses, the inability to measure marketing effectiveness accurately is not an abstract methodological inconvenience. It is a material business risk. Without reliable attribution, spending decisions tend to default to what is easy to measure rather than what is genuinely effective (Filippou et al., 2024). The consequence is not only financial inefficiency but a systematic undervaluation of marketing activities whose contributions manifest over longer time horizons and through indirect pathways. Most notably, brand-building investments and impression-first channels never generate a measurable click yet play a decisive role in shaping purchase intent(Morgan et al., 2002; Rust et al., 2004).

This thesis is motivated by these structural gaps. It begins from the practical reality that most small and mid-sized businesses possess only aggregated, channel-level data. From this constrained data environment, the thesis develops and validates a series of statistical modelling approaches that reveal the causal and associative relationships among marketing channels and outcomes. It offers a scalable, privacy-compliant, and practically actionable framework for marketing measurement.

1.2 Aims and objectives

The preceding section established that marketing measurement faces three intersecting challenges: the fragmented, multi-channel nature of modern digital consumer journeys; the temporal complexity introduced by advertising carryover and delayed effects; and the progressive erosion of user-level tracking data due to privacy regulation and platform-level restrictions. Taken together, these challenges leave a significant gap between what practitioners need to know the true incremental contribution of each marketing channel and what conventional attribution models are able to reliably deliver.

This thesis aims to address that gap through a programme of empirical and methodological research conducted in stages of increasing generality. Beginning from the established observation that digital marketing channels are interdependent (Naik and Raman, 2003; Kireyev et al., 2016), and extending through the practical constraints facing data-limited organisations (European Commission, 2023; Kamena, 2021), the thesis develops and validates a series of models that together constitute a coherent, privacy-compatible approach to marketing attribution. The overarching aim is to equip practitioners with rigorous, replicable tools for measuring the effectiveness and efficiency of their marketing investments.

The thesis is structured around four specific objectives, each corresponding to one of the published studies and each building on the evidence and limitations identified by its predecessors.

Objective 1:Establish the Cross-Channel Influence of Web Traffic on Direct Traffic Purchases

The first objective is to investigate whether, and to what extent, web traffic generated through individual digital marketing channels is associated with conversion activity attributed to the direct traffic source. Direct traffic is the single largest source of website visits (Bianchi, 2022*a*) and a meaningful contributor to e-commerce revenue (Kakalejčík et al., 2020).

Despite its importance, direct receives credit as a traffic source but managers cannot action on that insight. Whether direct traffic appears at the beginning, middle, or end of a conversion path, its presence in any attribution report offers little strategic direction. Managers cannot optimise a channel they cannot influence. Direct traffic is, by nature, a terminal channel. The point at which a user arrives and converts. Yet attribution frameworks treat it as a destination without interrogating its origin. The critical question is not how much credit direct deserves, but rather what drove the user there in the first place. Understanding the antecedents of direct traffic, rather than simply crediting it, is what enables managerial action and moves attribution from a descriptive exercise to a diagnostic one.

This exclusion is not merely a technical limitation; it represents a substantive measurement failure, because the direct traffic channel is understood to aggregate a heterogeneous mix of return visits, branded navigational searches, and conversions that were originally influenced by earlier, untracked touchpoints (Skow, 2023; Romero Leguina et al., 2020).

The consequence is that performance marketing channels are likely to receive artificially truncated attribution credit. Credit that ends at the last click recorded in the session, while the downstream conversion is assigned to a source that appears, misleadingly, to be channel-independent.

To address this, the first study analyses 89,394 purchase-level observations from a European e-commerce retailer and applies linear regression with multicollinearity and autocorrelation diagnostics to quantify the associations between channel-specific web traffic sessions and purchases completed via the direct traffic source (Filippou et al., 2024). The objective is specifically motivated by the call in the attribution literature for empirical evidence of cross-channel spillover effects (Kannan et al., 2016; Li and Kannan, 2014) and by the recognition that path-incomplete attribution substantially misrepresents channel value in practice (Danaher and van Heerde, 2018; Berman, 2018).

Objective 2: Examine the Effect of Performance Media Investment on Organic Branded Search Activity

The second objective is to determine whether investment in lower-funnel, performance-oriented digital advertising generates measurable brand awareness effects, as operationalised through organic branded keyword search volume. The separation of brand-building and performance objectives in digital media planning is deeply embedded in industry practice, yet has limited empirical support. Prior research has demonstrated that offline advertising: television in particular, drives online search activity (Joo et al., 2014, 2016; Chandrasekaran et al., 2018). Similarly, display advertising influences paid search (Kireyev et al., 2016; Zenetti et al., 2014). However, the question of whether performance campaigns, including paid social, paid search, display, and comparison channels, generate organic branded searches as a downstream effect has received substantially less attention. Branded search is itself a theoretically important metric: it reflects active consumer intent and brand recall (Srinivasan et al., 2005; Bernarto et al., 2020), and is understood to be a leading indicator of brand equity (Oh et al., 2020).

To address this objective, the second study applies ridge regression (Hoerl and Kennard, 1970) with adstock transformation (Gijsenberg et al., 2011) to weekly marketing investment and branded search data from a European e-commerce business operating across five countries. The use of adstock transformation is theoretically motivated by the recognition that advertising effects decay over time (Naik and Raman, 2003; Pauwels et al., 2004) and that models which ignore this temporal dimension will misestimate the relationship between media investment and brand-level outcomes. The specific choice of ridge regression responds to the documented multicollinearity among digital media channels (Polat and Gunay, 2015; Ma et al., 2019), and aligns with the broader literature on regression-based methodologies (Jin et al., 2017).

Objective 3: Develop a Causal Attribution Framework Operating on Aggregated Impression-Level Data

The third and central methodological objective of the thesis is to design, implement, and theoretically ground a novel marketing attribution framework that derives channel-level contribution estimates from aggregated, impression-level time-series data, without recourse to user identifiers, click logs, or individual conversion paths.

Existing data-driven attribution models whether path-based (Shao and Li, 2011; Kakalejčík et al., 2018), survival-theoretic (Zhang et al., 2014), Shapley-value-based (Zhao et al., 2018; Papapetrou et al., 2011), or deep learning-based (Arava et al., 2018; Ren et al., 2018) share a common dependency on user-level data. This dependency is increasingly difficult to satisfy. The GDPR (Li et al., 2019*a*; Zaeem and Barber, 2020), the CCPA (Bonta, 2022*a*), Apple's iOS 14 AppTrackingTransparency framework (Sokol and Zhu, 2021), and the deprecation of third-party cookies (Oksanen, 2022; Kamena, 2021) have collectively and materially restricted the availability of granular tracking data across the EU and beyond. The result is a structural misalignment between the data requirements of the dominant attribution paradigm and the data that is actually available to the majority of digital advertisers.

This objective directly responds to that misalignment. Drawing on the causal inference literature (Pearl, 2009*a*; Imbens and Rubin, 2015) and on recent advances in constraint-based temporal causal discovery (Runge et al., 2019; Shojaie and Fox, 2022), the third study develops the Causal-Driven Attribution (CDA) framework: a two-stage pipeline that first identifies the causal structure among marketing channels and conversion metrics using the PCMCI algorithm, and then estimates each channel's incremental causal effect through Monte Carlo simulation of Conditional Average Treatment Effects within a Structural Causal Model (Filippou, Quach, Lenghel, White and Jha, 2025). The framework is motivated theoretically by the reconceptualisation of attribution as a causal inference problem (Dalessandro et al., 2012; Hair Jr. and Sarstedt, 2021; Zhou et al., 2023) and practically by the need for an approach that remains operable under conditions of privacy-enforced data scarcity.

Objective 4: Evaluate Attribution Model Performance Through Simulation-Based Benchmarking

The fourth objective is to establish a rigorous, reproducible evaluation framework for the CDA model and to assess its performance across a range of causal graph complexities and conditions of structural uncertainty. A persistent limitation of the attribution modelling literature is the absence of ground-truth validation. In real-world settings, the true causal contribution of each channel is unobservable, making it impossible to assess model accuracy directly (Zhao, Mahboobi and Bagheri, 2019; Yao et al., 2022). Prior work has addressed this limitation through randomised field experiments (Johnson, 2023; Reiley et al., 2012), but such experiments are expensive, disruptive to normal business operations, and practically infeasible for the SME-scale organisations that are the primary audience of this thesis. An alternative, well-established approach is to generate synthetic data through a Structural Equation Model with predefined parameters (Iacobucci, 2009; Herman et al., 2025), so that ground-truth channel effects are known by design and model estimates can be evaluated against them.

To this end, the fourth objective involves the construction of a simulation environment that generates synthetic marketing datasets across five levels of causal graph complexity, with ground-truth channel contributions derived from the data-generating process. Model performance is evaluated using a multi-metric framework. It includes: causal effect estimation quality, relative RMSE, MAPE, Spearman rank correlation, causal graph recovery AUC, true and false positive rates, and structural graph distance metrics SHD and SID (Cheng et al., 2022; Tsamardinos et al., 2006). This evaluation design responds directly to calls for more transparent and comprehensive benchmarking standards in the attribution literature (Buhalis and Volchek, 2021; Gaur and Bharti, 2020) and provides a template that can be extended to assess competing attribution methods under controlled, comparable conditions.

The four objectives of the thesis are logically sequential and mutually reinforcing. The first two establish, on real data, that cross-channel effects are both directional and measurable, motivating the need for attribution frameworks sensitive to channel interdependencies. The third develops a novel framework capable of recovering these interdependencies under privacy-constrained conditions. The fourth validates the framework through rigorous simulation benchmarking, establishing the conditions under which it can be reliably applied. Together, they constitute a progressive empirical and methodological programme aimed at advancing the state of marketing attribution practice for the privacy-constrained digital environment.

1.3 Contributions

This thesis makes contributions across four interconnected dimensions: empirical, methodological, theoretical, and practical. Together, they advance the field of marketing attribution and measurement, with particular relevance to organisations operating in privacy-constrained and resource-limited environments.

1.3.1 Empirical Contributions

The thesis contributes three substantive empirical findings to the marketing analytics literature, each grounded in real business data or large-scale simulation.

First, it establishes that web traffic generated through specific digital marketing channels is positively associated with purchases made through the direct traffic source. The largest single source of website visits globally, yet one systematically excluded from conventional attribution credit (Bianchi, 2022*a*; Kakalejčík et al., 2020). Analysing 89,394 purchase-level observations from a European e-commerce retailer, the thesis demonstrates that sessions originating from Google paid search, price comparison engines, and email marketing collectively explain 61% of the variance in direct traffic purchases (Filippou et al., 2024). This provides empirical grounding for the widely held but under-evidenced hypothesis that indirect digital channels influence conversion outcomes through pathways that standard attribution systems cannot detect (Kannan et al., 2016; Berman, 2018).

Second, the thesis provides evidence that lower-funnel, performance-oriented advertising campaigns- can generate measurable brand awareness effects, as captured through organic branded keyword search activity (Filippou, Georgiadis and Jha, 2025). Using data from a European e-commerce business operating across five countries with an annual media budget exceeding €3 million, the analysis demonstrates that every major performance channel drives branded searches at a statistically significant level, with social media investment yielding the largest effect. This finding challenges the conventional separation of brand-building and performance objectives in digital advertising planning (Oh et al., 2020; Srinivasan et al., 2005) and offers direct empirical evidence that the indirect effects of performance campaigns are both real and measurable.

Third, at the aggregate simulation level, the thesis provides large-scale empirical evidence across 1,000 synthetic datasets. The datasets span five levels of causal graph complexity that channel-level attribution can be recovered from aggregated impression data with meaningful accuracy using a causal inference pipeline. Under the true causal graph, the framework achieves a relative RMSE of 9.50% and a Spearman rank correlation of 0.89 with ground-truth channel effects. Even when the causal graph must be inferred entirely from data, the framework achieves a relative RMSE of 24.23% while retaining a Spearman correlation of 0.53, demonstrating practically useful signal recovery under conditions of structural uncertainty (Filippou, Quach, Lenghel, White and Jha, 2025).

1.3.2 Methodological Contributions

The thesis introduces and validates several methodological innovations in the measurement of marketing effectiveness.

The first methodological contribution is the novel combination of ridge regression with adstock transformation for the purpose of measuring cross-channel marketing dynamics in a branded search context. While ridge regression has been applied to marketing data to address multicollinearity (Hoerl and Kennard, 1970; Polat and Gunay, 2015), and adstock transformation is well established in econometric modelling of advertising carryover effects (Gijsenberg et al., 2011; Naik and Raman, 2003), this thesis is the first to combine these two techniques to jointly address collinearity and temporal autocorrelation in this setting. The combination increases explanatory power substantially—raising R^2 from 0.658 to 0.768 relative to a standard linear model—while stabilising residual variance and providing a more faithful representation of advertising's fading effects over time (Filippou, Georgiadis and Jha, 2025).

The second methodological contribution is the development of the Causal-Driven Attribution (CDA) framework: a two-stage pipeline that integrates temporal causal discovery using the PCMCI algorithm (Runge et al., 2019) with causal effect estimation via a Structural Causal Model implemented in the DoWhy framework (Pearl, 2009*a*). Unlike all prior attribution methodologies, CDA operates exclusively on aggregated, impression-level time-series data, requiring no user identifiers, click sequences, or path-level behavioural records. The framework recovers directional causal relationships among marketing channels and quantifies each channel's incremental contribution to conversions through Monte Carlo estimation of Conditional Average Treatment Effects (Filippou, Quach, Lenghel, White and Jha, 2025). This positions CDA as the first impression-only causal attribution framework in the literature, directly addressing the structural data scarcity imposed by contemporary privacy regulation (Li et al., 2019*a*; Zaeem and Barber, 2020; Sokol and Zhu, 2021; Oksanen, 2022).

The third methodological contribution is a comprehensive simulation-based benchmarking framework for evaluating attribution models under controlled causal conditions. By generating synthetic datasets through a Structural Equation Modelling simulation with predefined causal graphs (Iacobucci, 2009; Herman et al., 2025), the thesis establishes a reproducible evaluation environment in which ground-truth channel effects are known by design. The framework assesses performance across six complementary metrics spanning both causal effect estimation quality—relative RMSE, MAPE, and Spearman rank correlation—and causal graph recovery accuracy—AUC, true and false positive rates, SHD, and SID (Cheng et al., 2022; Tsamardinos et al., 2006). This addresses a longstanding gap in the attribution literature, where the absence of observable ground truth in real-world data has historically precluded rigorous model validation (Yao et al., 2022; Zhao, Hua, Yan, Zhang, Xu and Yang, 2019).

1.3.3 Theoretical Contributions

The thesis advances attribution theory in several important respects.

It reconceptualises marketing attribution as a causal inference problem rather than a correlational measurement exercise (Hair Jr. and Sarstedt, 2021; Zhou et al., 2023; Dalessandro et al., 2012). By operationalising this shift through a unified pipeline of causal discovery and causal effect quantification, the thesis demonstrates that causal attribution can be achieved from aggregated data alone, extending prior theoretical arguments that had not been practically validated at this level of methodological integration. This represents a direct contribution to the growing body of work treating marketing measurement through the lens of the potential outcomes framework and structural causal modelling (Pearl, 2009*a*; Imbens and Rubin, 2015).

The thesis also contributes to the theory of cross-channel marketing dynamics by providing multi-study empirical evidence that the consumer journey contains substantial delayed and indirect effects that current measurement paradigms systematically underestimate. Performance campaigns produce brand awareness effects; paid channels influence direct traffic conversions; impression-first platforms shape downstream behaviour without generating traceable clicks. Collectively, these findings support a theoretical reframing of the consumer journey as a causally interconnected system (Naik and Raman, 2003; Pauwels et al., 2004; Schweidel, Bart, Inman, Stephen, Libai, Andrews et al., 2022) rather than a sequence of discrete, independently attributable touchpoints—consistent with calls in the literature for more structurally informed approaches to marketing measurement (Buhalis and Volchek, 2021; Gaur and Bharti, 2020).

Finally, the thesis addresses the privacy-attribution paradox: the growing divergence between the increasing analytical need for sophisticated attribution and the decreasing availability of the granular data on which conventional models depend. By demonstrating that meaningful attribution signal can be recovered from aggregated impression data using causal inference methods, the thesis establishes a theoretical proof of concept for a new generation of privacy-preserving attribution frameworks that do not require the erosion of user privacy as a precondition for effective marketing measurement.

1.3.4 Practical Contributions

Beyond academic contributions, this thesis produces several outputs of direct value to marketing practitioners, particularly within small and medium-sized enterprises.

The thesis provides a replicable modelling framework, combining adstock transformation with ridge regression. Practitioners can apply to their own channel data to measure the indirect branding effects of performance campaigns. The framework requires only standard aggregated analytics data, involves no proprietary infrastructure, and has been validated on real business data across multiple EU markets, making it immediately actionable for organisations without specialist data science resources (European Commission, 2023).

It provides the CDA framework as a scalable, computationally lightweight approach to privacy-compliant attribution that can be applied wherever aggregated impression and conversion time series are available. This is particularly relevant for industries subject to strict data protection regulation including healthcare, pharmaceuticals, telecommunications, and education (Onifade et al., 2021; Seth and Ramakrishnan, 2025)—as well as for any organisation facing the practical consequences of cookie deprecation and platform-level tracking restrictions (Kamena, 2021).

More broadly, the thesis equips marketing decision-makers with quantitative tools and interpretive frameworks that move beyond the surface-level metrics provided by standard digital analytics platforms, enabling more informed budget allocation, a more holistic understanding of channel interdependencies, and a more accurate assessment of the full-funnel contribution of each marketing investment. Given that direct experimental validation of channel contribution remains costly and operationally disruptive (Johnson, 2023), the observational and simulation-based methodologies developed here offer a practical alternative for organisations that cannot or do not run large-scale randomised media experiments.

1.4 Structure of the Thesis

The remainder of this thesis is organised into four chapters, each corresponding to a distinct phase of the research programme.

Chapter 2 presents the first empirical study, published in *Marketing Letters* (Filippou et al., 2024). The study investigates whether web traffic generated through specific digital marketing channels is associated with purchases completed via the direct traffic source, using 89,394 purchase-level observations from a European e-commerce retailer. The chapter details the conceptual framework, data collection procedures, regression modelling approach, and diagnostic tests for multicollinearity and autocorrelation, before reporting and interpreting the findings. It concludes by discussing the attribution implications of cross-channel influence on direct traffic and identifying limitations that motivate the subsequent study.

Chapter 3 presents the second empirical study, published in the *Journal of Marketing Analytics* (Filippou, Georgiadis and Jha, 2025). Building on the evidence in Chapter 3 that channel effects extend beyond their directly attributed conversions, this study examines whether lower-funnel performance media investments generate measurable brand awareness effects, as captured through organic branded keyword search activity. The chapter introduces the adstock transformation and ridge regression methodology, describes the data from a five-country European e-commerce business, and presents the modelling results. It then discusses the theoretical and practical implications of performance-to-brand spillover for media planning and attribution, and identifies the principal limitation that motivates the third study: the continued dependence on channel-level data only partially available under contemporary privacy constraints.

Chapter 4 presents the third and central methodological contribution of the thesis: the Causal-Driven Attribution (CDA) framework. This chapter describes the design and implementation of the two-stage pipeline combining PCMCI-based temporal causal discovery with Structural Causal Model-based effect estimation, and the simulation benchmarking study used to evaluate it across 1,000 synthetic datasets and five levels of causal graph complexity. The chapter details the data-generating process, the evaluation metrics, and the results under both oracle and estimated graph conditions, before situating the framework within the broader causal attribution literature (Dalessandro et al., 2012; Yao et al., 2022; Hair Jr. and Sarstedt, 2021) and discussing its practical scope and limitations.

Chapter 5 provides the concluding discussion of the thesis. It synthesises the findings across all three studies and reflects on how, taken together, they advance the empirical, methodological, theoretical, and practical contributions. The chapter discusses the broader implications of the thesis for marketing measurement practice in a privacy-constrained environment, considers the limitations of the work as a whole, and proposes directions for future research. It closes with a reflection on the thesis's overarching argument: that meaningful, causally grounded attribution is achievable without user-level data, and that this represents a viable and necessary path forward for the field.

The structure of this thesis is shown in Figure 1.1.

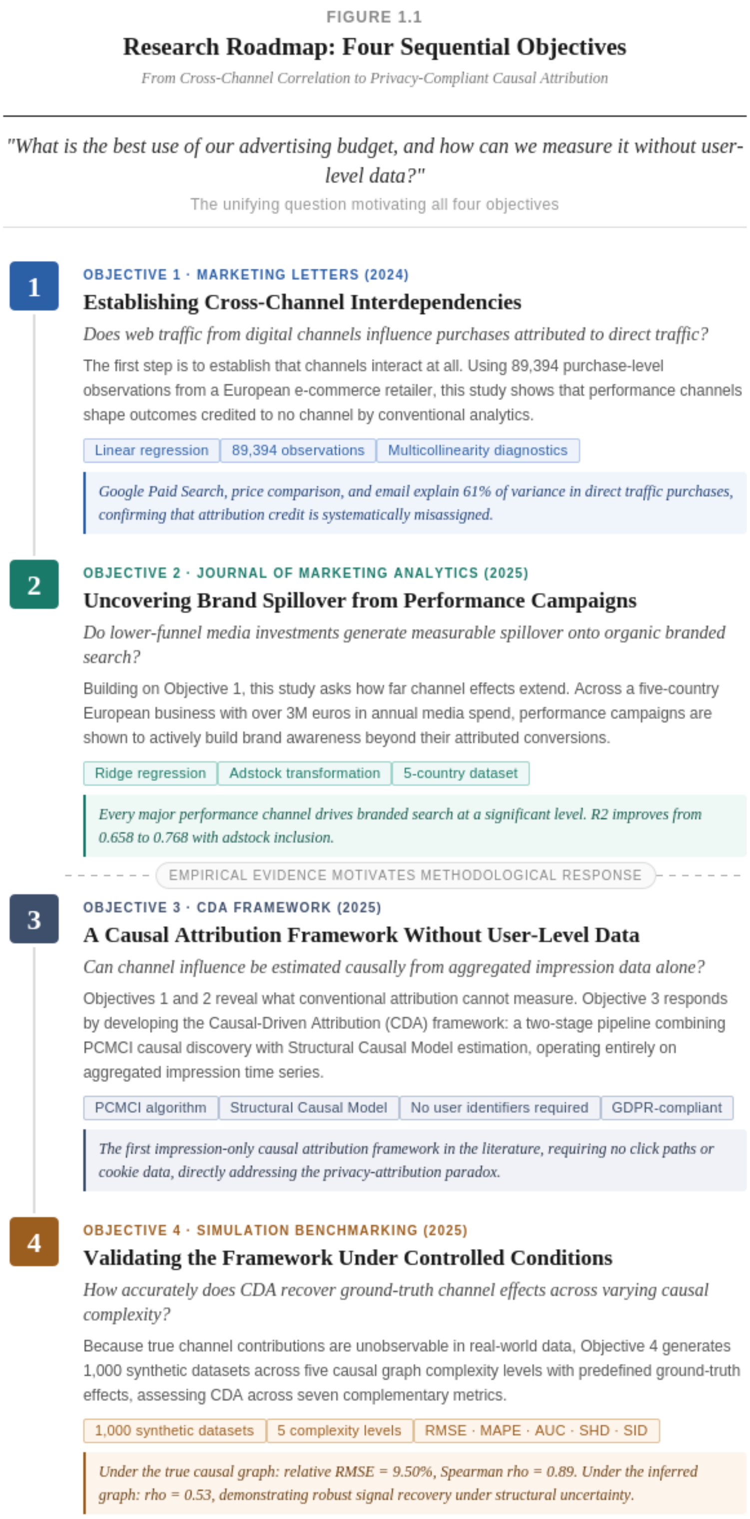


Figure 1.1: Research roadmap illustrating the four sequential objectives of the thesis, from establishing cross-channel interdependencies to validating a privacy-compliant causal attribution framework.

2

Establishing the link: Does web traffic from various marketing channels influence direct traffic source purchases?

Abstract

Marketing professionals and business owners strive to evaluate the effectiveness of their marketing investments. With multiple marketing channels at their disposal, understanding how these channels interact and influence each other is crucial. Digital analytics tools, such as Google Analytics, tend to measure the isolated success of each marketing channel. However, the intertwined effects and interdependencies between channels are often undervalued. This study, therefore, ventures into this territory. It focuses on the association between website traffic from various digital marketing channels and the purchases made by users visiting websites through direct traffic sources. We analysed 89,394 purchases from an e-commerce business in Europe. We conclude that three marketing channels can explain 61% of the variance. By shedding light on this overlooked aspect, we aim to guide advertisers toward a more holistic understanding of digital marketing channels.

2.1 Introduction

Global digital ad spending is predicted to exceed 626 billion USD in 2023 (Cramer-Flood, 2023), emphasising the role digital advertising holds for businesses. Advertisers have an array of marketing channels at their disposal, each with its distinct characteristics and advantages. Understanding the interactions and dynamics between these channels is important for effective marketing strategies.

While choosing marketing channels is critical, understanding the various traffic sources — how customers find and access websites — offers valuable insights (Wymbs, 2011). Users can visit websites through different sources, including paid search, organic search, social media, email, referral links, or direct entry of the URL. Each source reflects a different user behaviour and intent and plays a role in the customer's journey toward a purchase.

To navigate this complex environment, advertisers utilise different data strategies (Bhargava et al., 2020). Many of these strategies are provided by marketing technology, known as Martech. Martech consists of a broad array of software and tools that marketing professionals use to plan, execute, and measure marketing activities (Baltes, 2017; Brinker and Baldwin, 2020). However, the complexity of the customer journey and the potential for cross-channel interactions may present challenges for Martech tools in providing a complete picture of marketing effectiveness.

Using an observational method, we find a strong association between website visits from various channels and purchases from direct traffic sources. Although similar studies have used the same methodology, they have focused on different outcomes. Our study specifically investigates how sessions from multiple digital marketing channels influence purchases from direct traffic sources, which is an underexplored area in the existing literature.

This research seeks to contribute to the body of literature by addressing the understanding of digital marketing channel dynamics, particularly their influence on direct traffic. It provides practical insights and implications for marketers and businesses in the digital landscape. The following sections review relevant literature, present our hypotheses, describe our methodology, report results, and discuss their implications.

2.2 Theoretical background

The present section explores the literature surrounding marketing synergies, web attribution, and the consumer journey.

2.2.1 Marketing synergies and cross-media

Marketing synergies are the mutual enhancement of various marketing activities' impacts (Naik and Peters, 2009). Marketing synergies have been a research subject for decades, as understanding the effect of combined marketing channels can help advertisers plan more effective strategies (Edell and Keller, 1989; Naik and Raman, 2003).

The interactive effects of various media can enhance the overall performance of campaigns and improve key performance indicators (KPIs), ultimately influencing consumers' buying decisions (Chang and Thorson, 2004). For instance, Havlena et al. (2007) studied the isolated and synergistic effects of exposure frequency for TV, print, and internet advertising. They demonstrated that cross-media synergies could be quantified, providing a basis for understanding how different marketing channels interact and influence each other. These synergies can be particularly important for digital channels, where users frequently interact with multiple platforms and devices before making a purchase (Brookman et al., 2017).

2.2.2 Web attribution

Web attribution is a fundamental component of digital marketing, offering insights into which channels, touchpoints, and actions lead to a desired outcome on a website. Kitts et al. (2010) elaborates on how web attribution allows advertisers to understand the value of their marketing channels by assigning credit to each touchpoint in the customer journey. This method helps identify the most effective marketing channels, thereby improving decision-making and resource allocation.

More recent attribution models are data driven. These models leverage advanced analytics and machine learning algorithms to dynamically allocate credit based on patterns and trends in the data (Shao and Li, 2011). For instance, Anderl et al. (2016) propose a graph-based model that captures the sequence of customer interactions with different marketing channels. This model could provide more accurate attribution by considering the order and the content of customer interactions. Despite these advancements, the impact of indirect channels, specifically direct traffic, on purchase behaviour has received limited attention in the web attribution literature (Kannan et al., 2016).

2.2.3 Consumer journey

The consumer journey is the sequence of interactions or touchpoints a consumer has with a brand before reaching a specific outcome, such as purchasing (Hamilton and Price, 2019). This dynamic journey varies greatly among individuals and is influenced by a multitude of factors, including the advertising channels they encounter, the content they view, and their personal preferences and experiences (Malter et al., 2020). In the digital era, consumers interact with various online channels, such as social media, search engines, and email campaigns, each contributing to the overall consumer journey (Verhoef et al., 2015).

Website traffic plays a significant role in this modern consumer journey. A website acts as a central hub that provides information about products or services and as a platform for transaction completion. Various marketing channels funnel users to this hub, each contributing to the consumer journey in unique ways. Direct traffic, where users type in a web address or use a bookmark, is often an indicator of a more committed consumer, as it suggests prior knowledge or a direct intent to visit the website (Skow, 2023). However, the relationship between direct traffic and purchases from other marketing channels is complex and not yet fully explored in the literature (Kakalejčík et al., 2020).

2.3 Hypothesis development

In today's digital landscape, understanding digital marketing channels and their influences on consumer behaviour is pivotal. Some might argue that advertisements have minimal impact on conversions "but for" the users who were predisposed to purchase (Lambrecht and Tucker, 2013). However, the broader body of marketing literature suggests that advertising does indeed impact consumer behaviour, including online purchasing (Bleier and Eisenbeiss, 2015). Moreover, Lovett and Staelin (2016) provide evidence that paid media can drive direct traffic to a website, which, in turn, may lead to increased purchases.

This first hypothesis of our study is in contrast to previous studies, such as those by Plaza (2011), Omidvar et al. (2011), Milano et al. (2011), and Awichanirost and Phumchusri (2020), which primarily focused on different outcomes, such as website sessions or page views, rather than direct purchases. Our study focuses specifically on the relationship between website traffic from multiple marketing channels and direct traffic purchases. We propose the following hypothesis:

> **H1:** *Website sessions initiated through digital marketing channels are positively associated with purchases generated via direct traffic sources.*

While our first hypothesis involves the relationship between website visits and direct purchases, our second hypothesis is about the nature of this relationship. Existing research indicates that certain marketing channels have a more significant impact on consumer behaviour than others (Dalessandro et al., 2012; Kireyev et al., 2016). The digital marketing landscape is complex, with multiple channels each playing distinct roles in the consumer journey (Neslin and Shankar, 2009). As such, it is plausible that some channels have a more prominent association with direct traffic purchases. We therefore propose:

> **H2:** *Among the business channel mix, certain marketing channels (website sessions) have a higher association with purchases*

from direct traffic sources than others.

2.4 Methods

Researchers employ experimental research methods to establish and understand causal relationships. However, implementing such methodologies can sometimes be difficult or impossible (Johnson, 2023). The use of observational methods, while less able to establish causality, allows researchers to examine data in a less controlled yet more natural setting. When dealing with observational data, it is necessary to use analytical methods that can account for the inherent variability and complexity of real-world data (Asamoah, 2014).

2.4.1 Data

The data represent website performance from an e-commerce website operating in the European Union. The business requested to remain anonymous but shared access to their Google Analytics (G.A.) account. The product category sold is household goods, and the website operates in English.

Although the business received orders from 38 countries, we decided to focus on Europe (17 countries), as the number of purchases outside Europe was insignificant.

We gathered time-series data using Google Analytics for 180 days from August 2022 to January 2023. Following the cleaning process, the final report consists of a traffic report with sources, medium breakdown, and date. Table 2.1 presents the final dataset variables.

The final dataset consists of 7 columns. In the first one, we have Date. Then, we find the purchases from the direct traffic source (`purchases.direct`), which act as the response variable (y). The following five columns are our explanatory variables ($x_1, \ldots, x_5$): sessions from Google paid search (`google_paid_sessions`), sessions from price comparison engines (`price_comparison`), sessions from email marketing (`email_sessions`), sessions from organic search (`organic_sessions`), and sessions from social media (social_sessions).

Table 2.1: Dataset variables and descriptions

Variable	Type	Description
Date	Date	Daily observation date
purchases.direct	Response	Purchases via direct traffic source
google_paid_sessions	Explanatory	Sessions from Google paid search
price_comparison	Explanatory	Sessions from price comparison engines
email_sessions	Explanatory	Sessions from email marketing
organic_sessions	Explanatory	Sessions from organic search
social_sessions	Explanatory	Sessions from social media

2.4.2 Model

Digital marketing researchers can benefit from linear regression models to establish relationships among variables (Leeflang and Wittink, 2000). By employing a linear regression model, we can quantify the association between website sessions from different marketing channels (independent variables) and direct traffic purchases (dependent variable). The general form of the linear regression model is:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \varepsilon \tag{2.1}$$

where y is the response variable (direct purchases), $x_1, \ldots, x_p$ are the several channels performing as explanatory variables, β_j are the regression coefficients, and ε is the random error, which is often assumed to be normally distributed with mean zero and constant variance (James et al., 2013).

2.5 Results

Figure 2.1 shows the plots generated in R between the independent variables and several dependent variables. The plots showcase the observed relationship, and although further analysis is needed, there is a positive association between the independent and dependent variables.

In total, 16 models were created and evaluated. The best-performing model, Model 10, offers the highest R^2 (0.5344), and all coefficients are

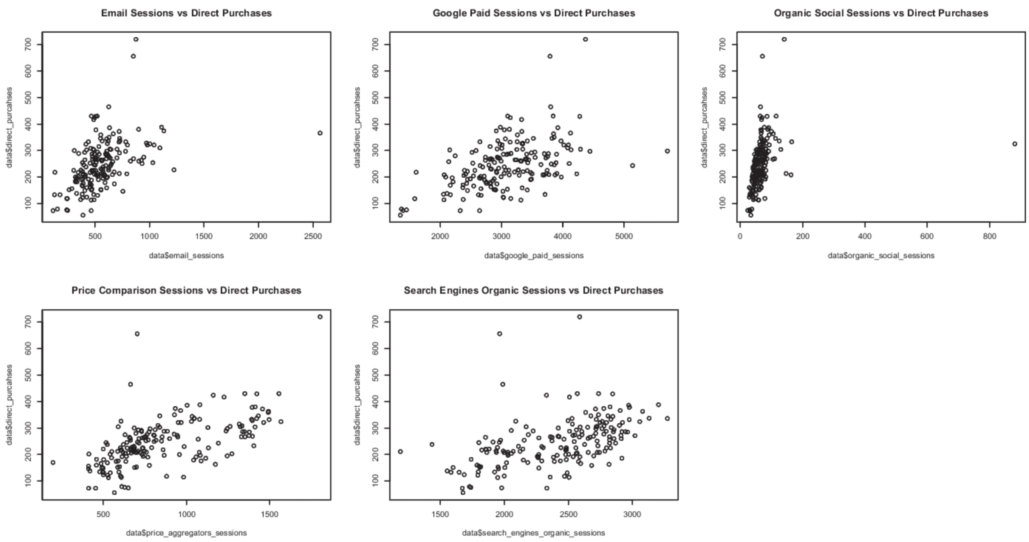


Figure 2.1: Dependent and independent variables (scatter plots)

statistically significant at the 0.001 level. As we will see in Section 2.5.1, after removing outliers the R^2 improves further to 0.613. Model 10 uses three explanatory variables: `google_paid_sessions`, `price_comparison`, and `email_sessions`. Table 2.2 presents the coefficient estimates.

The statistically significant relationship between direct purchases and our independent variables is still evident upon adjusting the initial model's variables by a scaling factor of 100 to facilitate interpretation. An increase of 100 sessions in `google_paid_sessions` results in 3.1 more direct purchases, a 100-unit increase in `price_comparison` results in 13.36 more direct purchases, and a 100-unit increase in `email_sessions` results in 6.7 more direct purchases.

Table 2.2: Model 10 — coefficient estimates (scaled by factor of 100)

Variable	**Estimate**	**Std. Error**	**Significance**
(Intercept)	4.08	1.14	< 0.001
`google_paid_sessions`	3.10	0.48	< 0.001
`price_comparison`	13.36	2.21	< 0.001
`email_sessions`	6.70	1.09	< 0.001
$R^2 = 0.613$ (after outlier removal); F-statistic: $p < 0.001$			

Leveraging the linear regression model and the adjusted proportions of daily average sessions from each channel, we can estimate the collective correlation of the predictors on direct purchases. Specifically, three channels —

google_paid_sessions, price_comparison, and email_sessions — collectively explain 61% of the variance in direct traffic purchases, supporting **H1** and **H2**.

These results could be attributed to a specific consumer behaviour pattern observed in digital marketing: users initially exposed to an advertisement via one of these three marketing channels often return later through a direct visit to complete their purchase (Bianchi, 2022*a*; Moe and Fader, 2004). This behaviour is consistent with a multi-step purchase process in which awareness is generated through paid or comparison channels and conversion occurs via a subsequent direct visit (Edelman, 2010; Google Analytics Help Center, 2023).

Indeed, data from the business's G.A. corroborate this narrative. They show that 8.45% of users visit the website for the first time directly. This implies that the vast majority of users (over 90%) must have reached the website through other channels prior to making a direct visit, reinforcing the cross-channel influence captured in our model (Bucklin and Sismeiro, 2009).

2.5.1 Model evaluation

When linear regression models are used, diagnostics for the model are essential to check the validity of the model. Therefore, standardised residuals and fitted values were checked to determine the model's assumptions (Sheather, 2009).

Figure 2.2 shows that there are outliers in the model. Therefore, we use standardised residuals (Sheather, 2009). Values outside the range $[-2, 2]$ are considered outliers and were therefore excluded. The new model reveals an improved R^2 of 0.613. Figure 2.3 presents the diagnostic plots after outlier removal.

For multicollinearity, we use the variance inflation factor (VIF) process. The results were not concerning, with VIF(google_paid_sessions) = 1.275 and with VIF(google_paid_sessions) = 1.275 and VIF(price_comparison) = 1.277, indicating no multicollinearity (Akinwande et al., 2015).

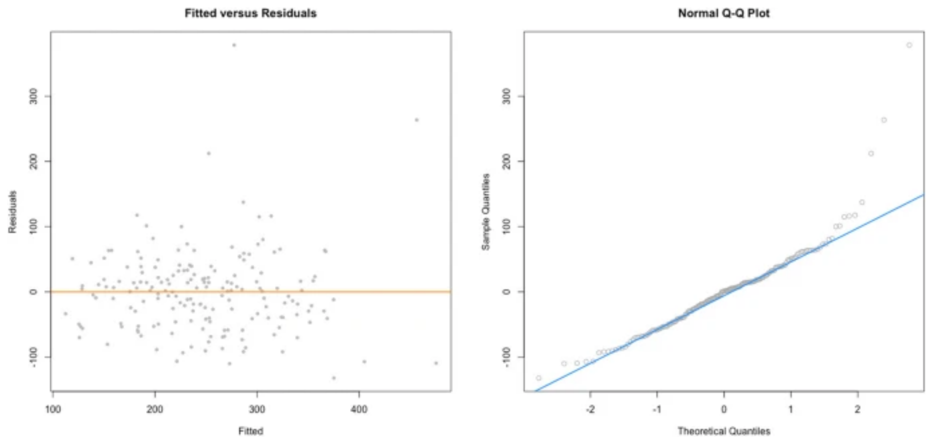


Figure 2.2: Fitted vs. Residuals and Normal Q–Q plot, Model 10

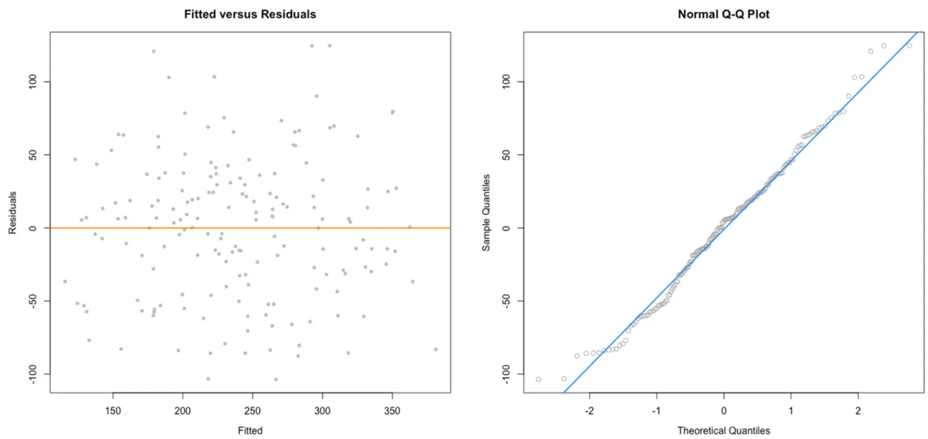


Figure 2.3: Fitted vs. Residuals and Normal Q–Q plot, Model 10 without outliers

Regarding autocorrelation, we performed the Durbin–Watson test (Dufour et al., 1980). The test did not reveal significant autocorrelation in the residuals (DW = 1.89), so no correction was required. This is consistent with Chatfield and Xing (2019), who note that daily retail data often show limited first-order autocorrelation once seasonal effects are controlled.

The next step is to understand whether the results are driven by seasonality. Seasonality presents a challenge in linear regression models because it introduces a systematic pattern of variation in the dependent variable (Hyndman and Athanasopoulos, 2018). We created a model using *Week Number* as the sole explanatory variable to assess whether sales are driven by time rather than the marketing channel sessions identified in Model 10. Week 1 represents the first week of the dataset, and Week 26 represents the last. This model reports a negative coefficient estimate for *Week Number* (-2.53), which is statistically significant, suggesting that direct purchases decreased slightly over time. However, with an R^2 of only 0.048, time explains far less variance than the three channel variables in Model 10, confirming that our findings are not simply an artefact of seasonality.

2.6 Discussion and conclusion

Our research contributes to three primary areas: academic literature, managerial practices, and marketing practitioners. From an academic perspective, our study bridges a gap in the current body of knowledge by investigating the relationship between various digital marketing channel sessions and direct traffic purchases. We provide empirical evidence that certain marketing channels are significantly associated with direct traffic purchases, adding a new dimension to the understanding of cross-channel dynamics in digital marketing (Dalessandro et al., 2012; Kireyev et al., 2016).

Our study outlines the importance of a more holistic approach to understanding web purchases from specific channels for managers. Li and Kannan (2014) claim that data-driven models can capture correlations between marketing touchpoints and purchase behaviour that are often missed by simpler attribution models. Our findings corroborate this, demonstrating that paid search, price comparison engines, and email marketing channels collectively explain a substantial proportion of the variance in direct traffic purchases. This provides managers with actionable insights for budget allocation across channels.

Finally, for marketing professionals, our work highlights an existing issue, revealing how specific digital channels might interrelate with each other, affecting the outcome of direct traffic purchases. As Xu et al. (2014) argue, the customer journey is not a single-step process, but a series of touchpoints leading to a purchase. Our study supports this view, showing that users engage with multiple channels before converting via the direct traffic source.

Our study has limitations inherent to using regression analysis for marketing research (Simon and Goes, 2013). Given the unique circumstances under which the relationships among variables occur, it is challenging to generalise our findings to different businesses, industries, or geographies. The data are drawn from a single European e-commerce business over a six-month period, and the results may not be representative of other sectors or market conditions. Additionally, while our model identifies associations, it does not establish causal relationships.

Future research should expand on our work. First, researchers should establish causal links between website sessions from various marketing channels and direct traffic purchases. This could be achieved through experimental designs or more advanced causal inference methods (Johnson, 2023). Second, future studies should investigate whether these associations hold across different business types, sectors, and geographic markets. Third, the role of unobserved confounders — such as brand equity, seasonal promotions, or concurrent offline campaigns — deserves further attention.

In summary, we discovered associations between the variables, and 61% of the variance was explained by three marketing channels (Google paid search, price comparison engines, and email marketing). The predictive power of these channels suggests that businesses should adopt a more integrated view of their marketing efforts, recognising that the influence of a channel extends beyond its directly attributed conversions. The findings also underscore the need for more sophisticated attribution models that can account for the complex interplay of channels in the modern consumer journey (Kannan et al., 2016; Neslin and Shankar, 2009).

3

Unveiling digital dynamics: Do digital media investments impact organic branded searches?

Abstract

This study addresses a fundamental challenge in digital advertising: how performance-focused campaigns, traditionally aimed at lower-funnel activity, may influence brand awareness. With the growth of online advertising, businesses increasingly rely on performance campaigns across social media, search engine marketing, and display networks to drive results. However, a notable phenomenon has emerged. Businesses may observe substantial branded search activity, suggesting lower-funnel campaigns could indirectly shape brand awareness. To investigate this relationship, data were analysed from an e-commerce business operating in five European countries with an annual multi-million-euro lower-funnel budget. We employed a ridge regression model with adstock values to account for temporal advertising effects and address common challenges of multicollinearity in multi-channel analysis. Our analysis reveals that lower-funnel campaigns significantly drive branded search activity, suggesting they play an unacknowledged role in building brand awareness. The incorporation of adstock values into ridge regression improves model performance, increasing the R^2 from 0.658 to 0.768 and lowering the root-mean-square error. This approach stabilises residual variance and provides a more accurate representation of advertising's fading effects over time. These findings offer a replicable framework for practitioners to measure the indirect branding effects of performance campaigns and optimise cross-channel investments.

Keywords: consumer behaviour; branded keyword searches; ridge regression; multi-channel marketing

3.1 Introduction

Branding serves as a cornerstone of marketing, enabling businesses to craft a unique identity that resonates with consumers and fosters recognition in competitive markets (Sarasvuo et al., 2022). Central to this process are distinct brand features, including names, logos, and design elements, which help establish a memorable presence (Manjunath et al., 2024). Regardless of whether a business operates in traditional or modern settings, a clear and consistent brand identity is crucial to differentiating itself and building consumer trust. To achieve this, companies employ diverse communication strategies aimed at capturing attention and encouraging loyalty (Xu et al., 2021).

With the rise of digital media, online branding has emerged as a dominant force, transforming how brands interact with consumers. Digital platforms now account for over 60% of global media investments, underlining the critical role of online channels in connecting with audiences (eMarketer, 2020). The internet's role in driving e-commerce has made strong online branding a necessity, as it significantly shapes consumer behaviour across demographics (Twenge et al., 2019). Research highlights this influence: approximately half the world's population has made online purchases (OECD, 2019), and 48% of consumers use the internet for shopping ideas and inspiration (Thygesen, 2022). This heightened engagement underscores the need for businesses to develop robust digital strategies. By leveraging platforms like Facebook, Instagram, and Google, which offer targeted ad placements, brands can enhance visibility, build loyalty, and achieve measurable performance outcomes (Sokolova and Titova, 2019; Ribeiro and Maués, 2023).

Digital platforms generally categorise their ad offerings based on 'Brand' (e.g. awareness or reach campaigns) and 'Performance' (e.g. conversion campaigns) objectives. For example, according to Facebook, 'brand campaigns' aim at maximising brand visibility, while 'performance campaigns' focus on driving purchases. Despite this industry structure, academic literature does not always mirror this approach, creating unique challenges for marketing practitioners when planning budget allocations.

The distinction between brand and performance marketing aligns with the consumer journey theory — a conceptual framework that maps the stages consumers experience as they engage with brands or seek fulfilling shopping experiences. This framework is critical for practitioners who leverage branding campaigns to attract consumers at the early stages of the journey, while performance campaigns are employed in later stages to drive conversions. Extensive literature supports this structured approach, demonstrating the impact of branding and performance strategies on consumer journey outcomes (Furquim et al., 2022; Tueanrat et al., 2021; Stankevich, 2017). Despite this, the consumer journey remains complex and often unpredictable, with individuals engaging with brands across various devices and touchpoints before making a purchase decision.

Tracking these multifaceted interactions is challenging, as highlighted by Vollrath and Villegas (2021). Rarely do consumers complete their journey in a single, direct interaction; rather, they tend to navigate through multiple touchpoints, devices, and platforms, often returning to a brand after several interactions. For instance, a consumer might begin by searching for a product online, then visit a brand's website via a sponsored link, only to revisit the site from another device or browser days later. The website re-visit could occur through various channels. This journey reflects how consumers weave through digital and offline experiences, creating a need for marketers to assign credit accurately to each touchpoint — an attribution process that is essential for assessing marketing effectiveness (Kannan et al., 2016).

For online businesses, the consumer journey is interpreted with the support of digital analytics systems. A widely used tool for identifying consumer journeys and patterns is attribution modelling, which is integrated into various modern web analytics platforms (McGuirk, 2023). According to Ben Mrad and Hnich (2024), attribution is "one of the most fundamental problems" modern marketers face. This challenge stems from the multitude of available traffic sources and the complexity of processing the resulting data. Consequently, the authors argue that businesses must deeply understand the relationships between various channels. These relationships, in some cases, can drive significant changes in budget allocation and, as a result, lead to major shifts in a business's strategic approach.

Attribution modelling relies on path data, which gives marketers valuable insights into users' behaviour. This data helps reveal patterns such as browsing habits, search preferences, and the steps users are likely to take before making a purchase decision (Hui et al., 2009).

Despite the wealth of research on digital attribution models, several important gaps persist, particularly for businesses running performance-driven activities. First, existing studies often overlook the broader context in which these businesses operate, thereby limiting the applicability of their findings to organisations that primarily focus on conversion and neglecting nuances that arise from industry, audience, or operational conditions. Second, most attribution frameworks rely heavily on path data — predominantly click-based information — leaving methods that do not depend on these sequential interactions underexplored. Third, although impressions can significantly influence consumer perception and intent (Qin et al., 2024), there has been relatively little development of models that incorporate impression-level data, potentially underestimating the role of brand visibility in driving performance. Finally, many attribution models presume that all business objectives centre on lower-funnel conversions, revealing a marked gap in methodologies that account for branding goals and top-of-funnel activities critical to long-term growth.

This study offers a novel perspective on assessing marketing effectiveness by moving beyond traditional click-based attribution models and embracing impression-level data as part of a privacy-first approach. Although it does not propose yet another attribution methodology, our method provides deeper insight into the complex interplay of marketing strategies, enabling performance advertisers to gauge the broader impact of their campaigns. By focusing on impression data, we highlight how performance-driven tactics can influence brand objectives, revealing that lower-funnel conversions are only part of a broader marketing success narrative. In doing so, we bridge a notable gap in existing research by demonstrating how even seemingly direct-response efforts may have downstream benefits for brand awareness and long-term customer engagement.

In order to answer our research questions, we develop a ridge regression model transformed with adstock values. The model uses observational data to delineate the relationships between variables in the consumer journey. To do this, a collaboration was established with a European e-commerce business operating in five countries and managing an advertising budget exceeding 3 million in 2023. Despite significant investments in performance-driven campaigns across channels like social media, search engine marketing, Google Display Network, and programmatic advertising, a substantial volume of branded keyword searches on search engines was observed by the business. This pattern gave rise to an essential question: how could brand awareness be so high when their focus had been exclusively on performance campaigns? The present study seeks to answer this by examining the indirect impacts of performance campaigns on brand visibility and awareness.

This model highlights the interplay between marketing channels, which is a topic of concern for many researchers (Naik and Peters, 2009). Moreover, it demonstrates the impact of these channels within the e-commerce sector. Importantly, our model integrates adstock values to account for the carryover effect, which reflects how advertising impact persists over time — a factor crucial in multi-channel contexts. Taking inspiration from Kannan et al. (2016), who emphasised the significance of the carryover effect in the context of attribution models, we adapted a ridge regression model to capture this extended influence across channels. The model performed significantly better than traditional linear regression, addressing autocorrelation challenges frequently encountered in empirical advertising studies.

This study makes several key contributions to both business analytics research and marketing practice. First, it establishes a robust way to demonstrate how paid advertisements, social media, and other online touchpoints interact to influence consumer behaviour in ways that standard single-channel or last-click models fail to capture. By incorporating adstock values, we not only address collinearity issues but also increase the accuracy (R^2) of our predictive models. Notably, we reveal a relationship between paid advertisements and brand awareness, demonstrated through branded search behaviour — even when the strategic intent did not explicitly target brand building. This discovery challenges common assumptions in traditional business analytics about the impacts and roles of different channels, showing that existing analytics approaches may undervalue certain tactics simply because their indirect effects are not adequately measured.

From a practical standpoint, the results provide marketing practitioners with a replicable framework for evaluating cross-channel effects in their own industries. By adopting similar modelling techniques, businesses can more confidently allocate budgets across digital channels and accurately gauge the long-term influence of their media spend. Moreover, our findings call into question the conventional, linear view of the consumer journey — highlighting that audience touchpoints are far more interconnected than typically assumed. This encourages business analytics researchers to rethink how they communicate and interpret multi-channel insights. Overall, our study underscores the need for deeper and more nuanced modelling techniques, ultimately guiding marketers towards data-driven strategies that better reflect the realities of consumer decision-making and channel interdependence.

The remainder of this chapter is structured as follows: first, the theoretical foundation is presented. This is followed by a description of the data collection and methodology. Next, the main findings are reported, and the chapter concludes with an analysis of the results and their implications for future research and practice.

3.2 Theoretical background

3.2.1 Consumer journey

The consumer journey metaphor illustrates the evolving relationship between consumers and brands, encompassing their shopping experiences and brand interactions. This journey spans multiple steps and engagements across diverse platforms and channels, often influenced by complex, abstract motivations (Hamilton and Price, 2019). Each stage reflects a unique blend of consumer intentions and brand messaging, shaping how consumers perceive and ultimately decide to interact with a brand.

Factors such as short- and long-term exposure to marketing messages can significantly influence consumers' journeys (Cambra-Fierro et al., 2021). In response, businesses approach the journey cautiously to foster an optimal customer experience (Dhebar, 2013). Customer journey mapping further enhances this approach by allowing businesses to visualise and understand the complete sequence of customer interactions, providing insights into how consumers experience their products or services (Lemon and Verhoef, 2016; Voorhees et al., 2017). For digital-first businesses, these interactions predominantly occur through digital touchpoints, presenting unique challenges and opportunities for engagement (Shavitt and Barnes, 2020).

In their comprehensive literature review on consumer decision journeys, Santos and Gonçalves (2021) observed that advancements in technology have fundamentally altered the consumer journey. Earlier, Verhoef et al. (2015) had highlighted a pivotal transition in retail from multi-channel to omnichannel approaches. This shift underscores how consumer experiences and decision-making processes have become more intricate, with a greater number of channels now shaping their journey. As such, consumers today navigate a complex web of channels, each contributing uniquely to their perceptions and ultimate purchasing decisions.

This research adds to the growing body of literature on consumer journeys by focusing on a specific pattern: users exposed to a marketing message who proceed to a branded keyword search after a few days, using search engines to return to the website. This delayed response to marketing represents a nuanced consumer journey that is not fully captured by existing attribution systems.

3.2.2 Attribution modelling

Attribution modelling plays a critical role in marketing by helping businesses allocate their budgets more effectively and gain insights into the customer journey. Recognised as a "thrust area for research in marketing" (Gaur and Bharti, 2020), the field has gained significant attention in the past decade. Attribution seeks to answer essential questions, such as identifying the source of a conversion, determining whether a single channel or a combination of channels contributed, and quantifying each channel's impact on the desired outcome (Ben Mrad and Hnich, 2024). E-commerce businesses, in particular, rely on business analytics tools to evaluate the performance of various touchpoints along the consumer journey (Bilgic and Duan, 2019). These tools are indispensable in understanding how exposure to advertisements influences consumer behaviour, revealing the contribution of different channels to specific actions (Kitts et al., 2010). However, standard attribution models, which depend on user-level data, face challenges in tracking consumer paths across platforms and devices, an issue further exacerbated by evolving privacy regulations and technical limitations (Brookman et al., 2017; Kamena, 2021).

To address these challenges, extensive research has been devoted to developing innovative methodologies for attribution. Markov Chain Monte Carlo (MCMC) models, which analyse path data to track the probability of transitioning between touchpoints until a purchase decision is made, have emerged as a prominent approach (Sandeep et al., 2021*a*; Poutanen, 2020; Kakalejčík et al., 2018). Another widely used method is the Shapley value model, which assigns value to each channel's contribution and addresses variability in attribution results, particularly for paid marketing channels (Gaur et al., 2024). These models offer significant advantages in understanding channel contributions but are increasingly limited by gaps in path data caused by stricter data privacy regulations and restrictions on third-party tracking (Romero Leguina et al., 2020). In response, researchers have proposed alternative solutions, such as logistic regression models for predictive accuracy (Shao and Li, 2011), causal methods leveraging observational data (Dalessandro et al., 2012), and multilevel models to analyse semi-continuous data over extended periods (Blozis, 2022). Despite these advancements, the quest for robust attribution methodologies continues, as businesses grapple with the complexities of measuring the effectiveness of marketing efforts in an evolving digital landscape.

3.2.3 Marketing synergies

Marketing synergies refer to the interplay between marketing initiatives that enhance the impact on consumer behaviour and brand engagement (Naik and Peters, 2009). Different marketing actions are interconnected, effectuating more streamlined and effective strategies (Havlena et al., 2007; Naik and Raman, 2003). Marketing synergies exist at multiple levels, encompassing both offline and online channels. Strategically focusing on marketing synergies in multi-channel systems can yield substantial benefits. By leveraging common operations and marketing across channels, businesses can create a unified customer experience and improve customer satisfaction and loyalty. For example, the convenience of online information searches combined with the tangible benefits of in-store purchases can attract a broader customer base (Gupta et al., 2004).

Understanding customer motivations is crucial in maximising these synergies. Some customers prefer online channels for information owing to easy accessibility, whereas risk-averse customers prefer offline channels for their perceived security (Ha and Perks, 2005). Others seek personal interactions available in physical stores, which online channels struggle to replicate despite integrating advancements such as avatars (Holzwarth et al., 2006).

According to Joo et al. (2014) and Kireyev et al. (2016), marketing channels have a positive influence on consumers' online branded keyword searches. They argue that advertisers should consider cross-media effects when planning advertising efforts. Joo et al. (2016) suggest that television advertising is positively related to branded search traffic. Similarly, Chandrasekaran et al. (2018) examine the interplay between offline TV advertising and online branded search behaviour. Alamsyah et al. (2014) develop a model to demonstrate that conversations on social media platforms affect brand awareness. From the various arguments in the literature, the following research question is posited. The first research question, R1, is on whether similar dynamics, as described by Joo et al. (2014), exist in an e-commerce environment:

> **R1:** *Do media investments from various marketing channels impact branded keyword searches for an e-commerce website?*

Drivas et al. (2021) claim that increased branded search traffic indicates a relationship between display advertising and brand awareness. However, they use secondary data sources, and their research objectives are different from those of this study. Using a similar approach, Devaux and Bomsel (2021) explore whether media investment in various channels positively impacts clicks. They find a clear association between clicks and brand awareness; however, they do not leverage media investment or branded searches. Kireyev et al. (2016) analyse whether display advertisements increase search clicks. Others examine the indirect effects of paid searches (Rutz et al., 2011) and other specific marketing channels. Zia and Rao (2019) find that advertisers allocate budgets across platforms that are asymmetric, either in terms of traffic or cost. To expand on the asymmetric approach to budget allocation, the second research question is:

> **R2:** *If a relationship exists between media spending on various marketing channels and branded keyword searches, which marketing channel has the biggest influence?*

This work is designed to contribute to managerial practices by providing actionable recommendations based on study findings. For instance, Filippou et al. (2024) illustrate their findings by scaling model outcomes to show the impact of a 1,000 investment. By using the model's coefficients, one can easily estimate the effect of a similar investment amount, making the implications more accessible for decision-makers. Similarly, the third research question is:

R3: *For each \$1,000 invested in any of the marketing channels of a business, how many branded searches, if any, would the business receive?*

3.3 Data

The data come from an e-commerce dropshipping business operating and fulfilling orders in five European Union countries. Specifically, 41.4% of total sales originated from Germany, making it the largest market. The Netherlands followed as the second-largest market with 24.6% of sales, while Belgium accounted for 12.2%, Spain for 10.8%, and Greece for 4.1%. Additionally, sales from other European countries where the business does not have an active domain or marketing campaigns were combined, representing 6.9% of total sales during the research period. The business is an online retailer operating in the homegoods industry, offering more than 200,000 products including home furniture, gardening, home appliances, and technology. Although different markets may require different strategic approaches (Chuang and Thomson, 2017), the business has adopted the same approach across all markets in which it operates.

As a business, they focus on price-sensitive purchases and therefore invest actively only in lower-funnel performance objectives across various marketing channels. They have return-on-investment (ROI) goals for all marketing activities and allocate budget whenever goals are not met.

Access was provided to their BigQuery for the analysis on the condition of remaining anonymous. Being a digital-first business, they do not have a physical presence and invest in several digital marketing channels. The dataset contains 360 days of time-series data for March 2023–2024. The business invested roughly 3 million in 2023 in various digital marketing channels. Post-cleaning, the data included daily investments in search engine marketing (SEM), display, programmatic advertising, and aggregated lower-volume channels (social media including Facebook, Pinterest, and Instagram), labelled as 'others' (x_j). SEM represents 63.04% of the total investment, display advertising 18.06%, programmatic 15.20%, and other channels 3.7% for the period of the study.

Finally, daily branded keyword searches were taken as the response variable y. Any query comprising the brand name as part of an organic branded search was included in the data for analysis.

3.4 Methodology

Randomised control trials are the gold standard for testing hypotheses in the industry, allowing researchers to identify and understand causal connections through experimental research techniques. However, applying these methods often presents challenges (Johnson, 2023). A non-experimental approach is used here, as media spending could not be allocated and an experimental environment could not be created.

Digital marketing studies often use linear regression models to determine the correlations between variables (Leeflang and Wittink, 2000). These models measure the degree to which factors such as media spending across marketing channels correlate with organic branded search activity. Following Zamil et al. (2021) and Filippou et al. (2024), a multiple linear regression model is deployed initially.

Businesses generally maintain consistent proportional allocations across marketing channels, even when total spending changes. For example, if a company allocates 10% of its budget to social media, this percentage remains constant regardless of the investment size. Consequently, media investments are inherently correlated as predictors. To mitigate this problem, ridge regression is employed.

3.4.1 Ridge regression

Hoerl and Kennard (1970) demonstrate that regression coefficients can be excessively large or incorrectly assigned when the prediction vectors are non-orthogonal. A typical model for multiple linear regression is:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{3.1}$$

where $E(\boldsymbol{\varepsilon}) = \mathbf{0}$, $E(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}') = \sigma^2\mathbf{I}_n$, and the design matrix $\mathbf{X}$ is of dimension $n \times p$, with n being the data size and p the number of predictors. The variables in $\mathbf{X}$ are standardised such that $\mathbf{X}'\mathbf{X}$ forms a correlation matrix. The vector $\hat{\boldsymbol{\beta}} = \mathbf{X}'\mathbf{y}$ consists of the correlation coefficients between the response variable and each predictor.

Ridge regression modifies typical linear regression by adding a penalty to the size of the coefficient. Specifically, the regression coefficients are estimated as:

$$\hat{\boldsymbol{\beta}}(k) = (\mathbf{X}'\mathbf{X} + k\mathbf{I})^{-1}\mathbf{X}'\mathbf{y}, \quad k \geq 0, \tag{3.2}$$

where k is known as the ridge parameter, controlling the amount of shrinkage applied to the coefficients. When $k = 0$, the ridge estimator is equivalent to ordinary least squares, which is unbiased but has large variance when predictors are highly collinear. As k increases, the estimator becomes more biased but typically has lower variance. The key to ridge regression is the ridge trace — a plot of the estimated coefficients as a function of k — by examining which the value that stabilises the coefficient estimates can be identified.

Ridge regression algorithms are used in the media industry (Ma et al., 2019) and more specifically to describe consumer-journey-related challenges (Sun et al., 2022). However, this is the first time ridge regression has been used to explain the relationships among variables such as branded keyword searches and media investment. Moreover, Joo et al. (2014) and Kireyev et al. (2016) find highly correlated variables in similar datasets but do not leverage ridge regression, a method that effectively addresses collinearity.

3.4.2 Adstock transformation

In this study, an innovative approach around model transformation is adopted by incorporating adstock values for each variable. Adstock is the cumulative value of a brand's advertising impact considering the effect over time (Gijsenberg et al., 2011), and it is "central to the econometric modelling of advertising effects over time" (Ephron and McDonald, 2002). This approach, which incorporates a time lag into the model, is frequently used (Chandrasekaran et al., 2018; Joo et al., 2014; Kireyev et al., 2016). For an adstock model:

$$S_t = \theta A_t + (1 - \theta) S_{t-1}, \tag{3.3}$$

where S_t is the adstock-adjusted value of advertising impact at time t, A_t is the advertising expenditure at time t, θ is the decay rate (or retention rate), a parameter between 0 and 1 representing the fraction of the advertising impact retained from one period to the next, and S_{t-1} is the adstock value from the previous period (Joseph, 2006).

3.4.3 Model estimation

To finalise the model, the adstock-transformed values for each media channel are computed first:

$$S_{j,t} = \theta_j x_{j,t} + (1 - \theta_j) S_{j,t-1}, \tag{3.4}$$

where $S_{j,t}$ is the adstock value for channel j at time t, $x_{j,t}$ is the expenditure on channel j at time t, and θ_j is the decay rate for channel j, constrained to $0 < \theta_j < 1$. Once the adstock values $S_{j,t}$ are calculated, they are used as predictors in the ridge regression model:

$$Y_t = \beta_0 + \beta_1 S_{1,t} + \beta_2 S_{2,t} + \beta_3 S_{3,t} + \beta_4 S_{4,t} + \varepsilon_t, \tag{3.5}$$

where Y_t is the organic branded searches; $S_{1,t}$, $S_{2,t}$, $S_{3,t}$, and $S_{4,t}$ are the adstock-adjusted spend for SEM, display advertising, programmatic advertising, and aggregated lower-volume channels, respectively; $\beta_0, \beta_1, \beta_2, \beta_3$, and β_4 are the coefficients to be estimated; and ε_t is the error term at time t.

Next, the value for ridge parameter λ is chosen, which balances bias and variance. The model is fitted using ridge regression to minimise the penalised residual sum of squares (PRSS):

$$\text{PRSS} = \sum_{t=1}^{n} (Y_t - \beta_0 - \beta_1 S_{1,t} - \beta_2 S_{2,t} - \beta_3 S_{3,t} - \beta_4 S_{4,t})^2 + \lambda\left(\beta_1^2 + \beta_2^2 + \beta_3^2 + \beta_4^2\right), \tag{3.6}$$

where β_0 is the intercept capturing the baseline level of organic branded searches when media spend is zero, β_j is the coefficient for each media channel's adstock-transformed spend, $S_{j,t}$ is the adstock-adjusted spend for each media channel at time t, and λ is the regularisation parameter controlling the extent of shrinkage applied to the coefficients. The summation ranges from $t = 1$ to n, where n is the total number of observations. The PRSS balances model fit with a penalty term to prevent overfitting.

3.5 Results

The first version is a linear regression model. An R^2 of 0.666 and statistically significant variables are recorded. Specifically, the SEM coefficient (x_1) is 0.274, display (x_2) is 1.778, programmatic (x_3) is 0.359, and others (x_4) is 4.612, with an intercept of 160.644.

The challenge arises from Figures 3.1 and 3.2, which show the fitted-versus-residuals and normal Q–Q plots, respectively. The clear pattern in Figure 3.1 and the outliers in Figure 3.2 impel further exploration of model transformations (Sheather, 2009) without succeeding in resolving this issue.

In the second version, adstock values are incorporated into the linear model, a common approach in the field (Jin et al., 2017). Applying the correct adstock values in modelling is complex (Gijsenberg et al., 2011). The business data in this study indicate that approximately 95% of purchases occur within the first 2 days of the initial visit. Given this rapid conversion timeline, a lower adstock rate reflects the short-lived impact of initial engagements. Therefore, a geometric adstock value of $\theta = 0.1$ is applied for all channels.

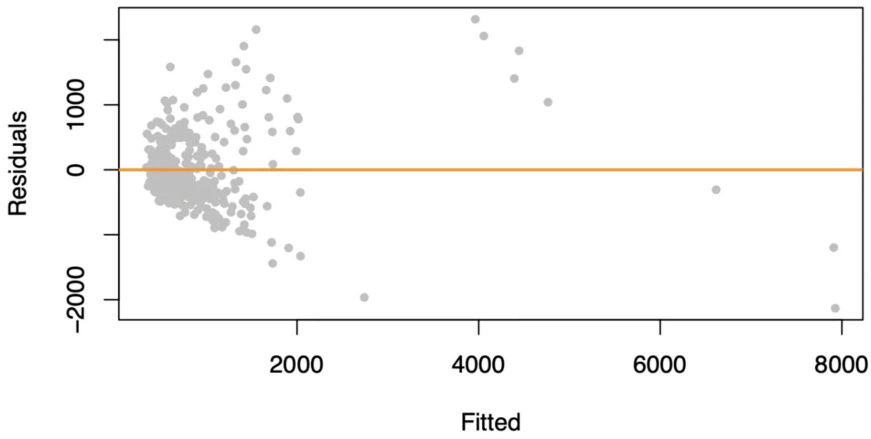


Figure 3.1: Fitted vs. Residuals plot, baseline linear regression model

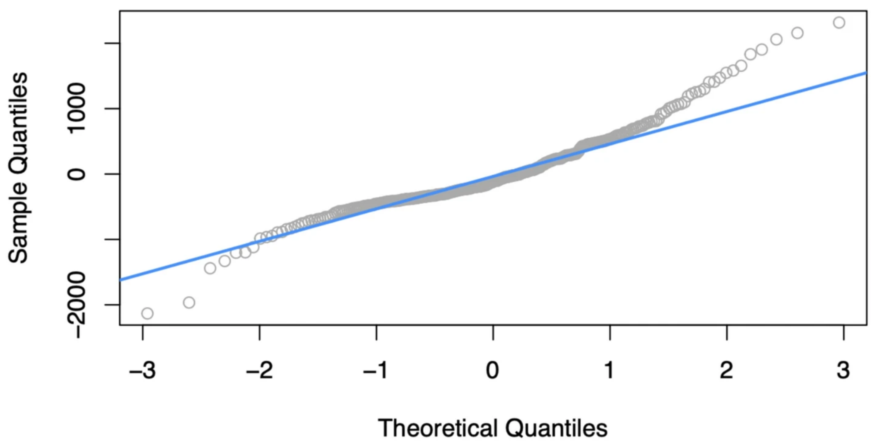


Figure 3.2: Normal Q–Q plot, baseline linear regression model

An R^2 of 0.675 is obtained, a slight improvement on the initial model, and all predictors remain statistically significant. Since all coefficients remain similar, the problem of model fit must be addressed. A probe for collinearity reveals a high correlation of 0.89 between the SEM and display variables (Chen, 2016). Collinearity undermines the reliability of a linear model, particularly when predictors such as marketing channel expenditures tend to fluctuate simultaneously. High correlations among the variables can induce inflated standard errors, making it difficult to discern the true effects of each predictor. Ridge regression manages multicollinearity by introducing a penalty term that shrinks the coefficients (McDonald, 2010); thus, it is a promising approach.

The result of the ridge regression model with $\lambda = 191.048$ (Figure 3.3) yields $R^2 = 0.658$. The SEM coefficient (x_1) is 0.278, display (x_2) is 1.390, programmatic (x_3) is 0.355, and others (x_4) is 5.033, with an intercept of 223.483. The ridge regression model does not improve the fitted-versus-residuals pattern.

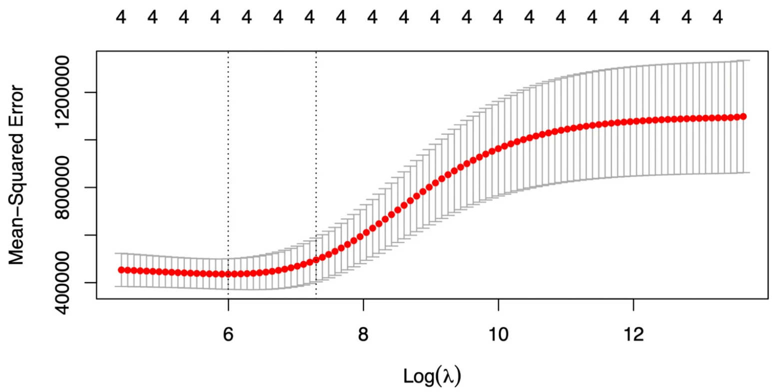


Figure 3.3: Ridge trace plot, $\lambda = 191.048$

As the ridge regression model does not improve the fitted-versus-residuals pattern, the linear model is compared with the ridge model based on the root-mean-square error (RMSE). RMSE quantifies the average magnitude of the prediction errors in a regression model and is a critical indicator of model performance by measuring the difference between the predicted and observed values (Hodson, 2022). Other researchers have used this comparison across various fields (Polat and Gunay, 2015). After splitting the data (80/20), the results indicate that the ridge model fits better, with a smaller RMSE of 552.6 compared with the linear model's 640.3. Therefore, the ridge regression model is iterated further.

In the next step, adstock values are applied to the ridge model to capture the diminishing influence of advertising over time. Incorporating adstock values allows for reflecting the real-world fading effect of advertisements, thus providing a more accurate and temporally sensitive model (Pauwels et al., 2004). This is the first study to incorporate adstock values into a ridge regression model to address a business challenge of this kind.

The results of the final model improve substantially, with $R^2 = 0.768$. Table 3.1 presents the coefficient estimates for all models. All predictors in the final model are statistically significant. As shown in Figure 3.4, the pattern in the standardised residuals is eliminated. The inclusion of adstock values in the ridge regression model significantly stabilises the variance among the residuals, yielding a more robust and reliable model.

Table 3.1: Coefficient estimates across all model specifications

	Linear	**Linear + Adstock**	**Ridge**	**Ridge + Adstock**
R^2	0.666	0.675	0.658	**0.768**
RMSE	640.3	—	552.6	**552.6**
Intercept	160.644	—	223.483	202.191
x_1 (SEM)	0.274	—	0.278	0.243
x_2 (Display)	1.778	—	1.390	1.340
x_3 (Programmatic)	0.359	—	0.355	0.373
x_4 (Other)	4.612	—	5.033	5.407

All predictors statistically significant in final model. $\theta = 0.1$ applied to all channels.

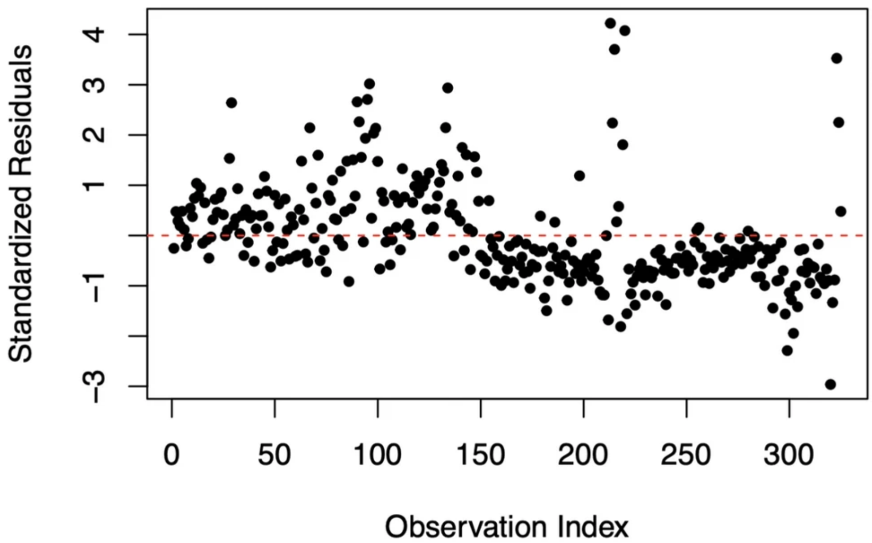


Figure 3.4: Fitted vs. Residuals plot, final ridge regression model with adstock

3.6 Discussion and conclusion

This study examines the complex relationship between digital media investments and online organic branded keyword searches in the e-commerce context using ridge regression with adstock values. This approach addresses the prevalent issue of collinearity in marketing data and significantly improves model robustness and accuracy.

Addressing R1, the final model enhances understanding of the dynamic interplay between media channels. The model achieves an R^2 of 0.768, indicating a strong fit that explains 76.8% of the variability in branded keyword volumes by media investment.

Addressing R2, the coefficients derived from the ridge regression model provide clear insights. The analysis indicates that each marketing channel has a quantifiable effect on branded keyword searches. The most influential channel is 'others', representing aggregated investments in social media, with a coefficient of 5.407. Display advertising is the second-largest driver of organic branded keyword searches with a coefficient of 1.340. Programmatic advertising ranks third with a coefficient of 0.373, followed by SEM with a coefficient of 0.243. The intercept is 202.191.

Addressing R3, the impact of each $1,000 invested was quantified using the model coefficients. For every $1,000 invested in social media, 5,407 branded searches are generated, making social media the most influential marketing channel for driving brand awareness. A further $1,000 invested in display advertising effectuates an additional 1,340 organic branded searches. Programmatic advertising yields 373 additional searches, and SEM results in 243 additional branded searches per $1,000 invested.

Initial data exploration confirmed that 'organic' channels ranked highest, with branded keywords prominently driving consumer interactions. This observation reflects a common consumer journey challenge, as it suggests that users often engage with a brand organically even after exposure to various paid media channels. This finding presents a fundamental challenge in understanding the path to purchase, as the channels that drive initial engagement may differ from those associated with conversions. Previous studies by Huang and Lin (2022) emphasise the role branded keywords play in shaping the consumer journey, a notion corroborated by Bianchi (2022*b*), who found that 42% of global searches are branded, indicating a strong brand-oriented mindset among consumers at the search stage. Consumers often arrive at search engines with specific brands in mind, suggesting that organic channels are crucial in the final stages of the journey.

Furthermore, the analysis details the reduction in prediction error using ridge regression compared with a traditional linear model. The RMSE of the ridge model is 552.6, compared with 640.3 of the linear model, demonstrating a more accurate and reliable fit.

3.6.1 Managerial implications

Brand awareness is a critical component of marketing strategy (Bernarto et al., 2020), yet efforts to predict and quantify it remain challenging (Srinivasan et al., 2005). Existing attribution methodologies often complicate budget allocation decisions (Danaher and van Heerde, 2018), leaving marketers searching for more practical and effective approaches. This study contributes to managerial knowledge by addressing these gaps, offering actionable insights into how media investments influence branded searches and, consequently, brand awareness.

First, we demonstrate that certain marketing strategies may have unintended side effects. Specifically, our findings reveal that lower-funnel performance campaigns, which may initially appear inefficient, generate branded searches and thus contribute positively to brand awareness. This insight challenges traditional perceptions of campaign performance and encourages managers to consider broader implications beyond immediate metrics.

Second, this study introduces a simple, replicable model that managers can apply to their own businesses to predict increases in branded searches. By overcoming issues such as multicollinearity, our approach ensures robust and reliable results, equipping practitioners with a practical tool to better understand the complex interplay of marketing channels and their contributions to brand-building efforts.

Third, we challenge modern managers to move beyond the surface-level data provided by digital analytics platforms. While clicks and path data are valuable, they fail to capture the full impact of marketing strategies. By incorporating advanced modelling techniques, practitioners can uncover deeper insights into the effects of their campaigns, enabling more strategic and effective decision-making.

Finally, while this study is not the first to highlight the importance of impressions in marketing analytics, it reinforces the need to prioritise impressions alongside clicks and path data (Sirdeshmukh et al., 2018). By addressing the challenges of omnichannel strategies (Cui et al., 2021) and the cannibalisation of channels (Shankar and Kushwaha, 2021), this study provides managers with the tools and insights necessary to navigate the complexities of modern marketing.

3.6.2 Research contributions

This research contributes to the literature by exploring the consumer journey from web advertising exposure to branded searches, focusing on advertisers investing in various media channels and identifying a relationship between media investments and branded searches. The study also highlights the importance of marketing synergies and relevant work related to the interplay between marketing channels.

From an academic perspective, this is the first study to combine ridge regression and adstock values to address the challenges of the consumer journey through a simple and robust method when collinearity causes problems. For instance, if Drivas et al. (2021) and Devaux and Bomsel (2021) had used a similar approach, the collinearity issues in their results could have been addressed. Similarly, our novel approach overcomes patterns with residuals indicating autocorrelation (Chen, 2016). By replicating the approach used in this study and diverging from traditional linear regression models, better model performance with lower RMSE and higher R^2 values can be obtained while maintaining model validity.

3.6.3 Limitations and future directions

The specific limitations associated with regression analysis in marketing research are recognised in this study (Simon and Goes, 2013). The conditions under which variable relationships manifest pose difficulties in generating theories from correlational data. Moreover, the potential challenge of directionality underlies the findings (Asamoah, 2014). However, the data indicate that most individuals accessing a site through branded search terms are doing so for the first time (over 90%), affirming their status as independent variables. Additionally, the potential impact of extraneous variables on the results must be considered (Lovett and Staelin, 2016). As businesses control changes in media spending, the impact of extraneous variables may be reduced.

In ridge regression, determining the specific value of k for collinear explanatory variables involves both art and science. Most studies focus on identifying an optimal ridge parameter k that ensures the ridge estimator has a lower mean squared error than the least squares estimator. Although many effective methods can be used to determine the optimal k value, no single method has been universally adopted. Finding ridge estimates that offer more utility than least squares estimates should therefore be a practical aim of future studies (McDonald, 2009).

A significant challenge also lies in accurately measuring adstock effects, complicating the assessment of the precise impact of advertising on sales over time (Gijsenberg et al., 2011). With over 95% of purchases occurring within the first 2 days, the adstock variable was set to 0.1 in this study, indicating a short consumer cycle.

Future research should explore whether similar relationships exist between other channels, potentially expanding online and offline media mixes. Furthermore, researchers can replicate this approach using different media types, such as email or influencer marketing. An extended dataset from similar or different industries can provide further insights. Additionally, an experimental approach, such as geo-experiments, can offer additional causal evidence.

4

Causal-driven attribution (CDA): Estimating channel influence without user-level data.

Abstract

Marketing attribution is essential for data-driven budget allocation. Yet conventional approaches rely on user-level path data that are increasingly unavailable. We propose a Causal-Driven Attribution (CDA) framework that enables marketers to quantify channel-level contributions to conversions using only aggregated impression data. The framework first uses a temporal causal discovery algorithm (PCMCI) to identify which marketing channels influence one another and in what direction, then applies Structural Causal Models (SCMs) to quantify how much each channel individually and in combination, contributes to conversions. We validate CDA on large-scale synthetic data that mimic realistic multi-channel marketing scenarios. When the true channel relationships are known, the framework achieves an average relative RMSE of 9.50% when these relationships must be learned from data, the error rises to 24.23% demonstrating that CDA performs well overall, and particularly so when the underlying interaction structure is correctly specified. Beyond methodological contributions, CDA offers practitioners an interpretable, privacy-compliant attribution tool that captures cross-channel synergies, supports strategic media planning, and remains operational as third-party tracking continues to erode. The framework provides a scalable alternative to path-based models, aligning marketing analytics with the evolving data privacy landscape.

Keywords: Marketing attribution, multi-channel marketing, causal inference, PCMCI, structural causal model, digital advertising

4.1 Introduction

> "When budgets are tight, marketers need to justify every penny of their spend. But without full-funnel attribution, spending decisions are often based on what is easy to measure rather than what is actually working." — *MarketingWeek, 2025*[1]

As consumers increasingly engage with marketing stimuli across multiple channels, more sophisticated frameworks are required to evaluate whether investments across these channels actually generate the desired outcomes (Berman, 2018). This motivation gives rise to the concept of marketing effectiveness, which represents the degree to which marketing investments drive the desired results (Solcansky and Simberova, 2010).

However, measuring marketing effectiveness has become increasingly challenging in the digital era. Many researchers have employed diverse methodologies to evaluate it, including experimentation, qualitative and quantitative observational analyses, econometric modelling, and machine learning approaches. Among these, statistical models have become particularly prominent as rigorous tools for assessing marketing effectiveness (Zenetti et al., 2014; LaBrecque et al., 2024). Organizations increasingly rely on such models to gain actionable insights guiding strategic decisions, particularly regarding advertising budget allocation under uncertainty (Qin et al., 2024).

One of the most widely used statistical approaches for defining marketing effectiveness is marketing attribution (Buhalis and Volchek, 2021). It has received growing attention as both a methodological and theoretical challenge. Attribution refers to assigning credit to different marketing touchpoints influencing customer purchase paths. Drawing from attribution theory in psychology (Kelley, 1967; Heider, 2013), this process involves determining causal relationships between marketing stimuli and consumer responses.

The proliferation of entertainment and advertising platforms has added further complexity to understanding consumer behaviour (Schweidel, Bart,

[1] https://www.marketingweek.com/full-funnel-attribution/

Inman, Stephen, Libai, Andrews, Rosario, Chae, Chen, Kupor et al., 2022). Simultaneously, changes in digital privacy, particularly Apple's iOS14 updates (Sokol and Zhu, 2021), the European Union's General Data Protection Regulation (GDPR) (Li et al., 2019*b*), and the California Consumer Privacy Act (CCPA) (Bonta, 2022*b*), have significantly restricted access to individual level data. In a cross-channel interaction environment, where users can move freely across touchpoints and marketing messages appear across numerous platforms and devices, these updates create significant difficulties for accurate measurement due to the lack of data transparency throughout the consumer journey (Cui et al., 2021).

Limited data availability creates what we term the "privacy-attribution paradox" where the need for sophisticated attribution increases as data accessibility decreases. This paradox is compounded by both technical and conceptual limitations that existing literature treats separately but are fundamentally interconnected. Technical limitations include privacy restrictions, limited data access, and systematic mislabelling of traffic sources (Starov et al., 2018; Filippou et al., 2024). Conceptual limitations arise from the model's inability to provide transparent interpretability. The results often fail to reveal the precise insights researchers seek. For example, we may desire to understand the causal relationships between marketing channels rather than the correlations among them.

Furthermore, the emergence of view-first (or impression-first) entertainment platforms exemplifies these interconnected challenges. Reiley et al. (2012) found that standard click-based models can misattribute the advertising impact by assigning credit to clicks from users who were likely to convert regardless of exposure to the advertising. Users engage deeply with content without necessarily clicking or transitioning to websites, creating influential but unmeasured interactions. Academic research demonstrates that impressions without website visits can generate sales (Ghose and Todri-Adamopoulos, 2016), while industry studies show significant impact of views and impressions on conversion paths without clicks (TikTok for Business

2024[2]; Snapchat for Business 2024[3]).

Marketing practitioners cannot analyse the full user journey since each ad platform tracks its own touchpoints. This creates structural opacity in digital attribution (Romero Leguina et al., 2020), leading to biased or incomplete attribution outcomes. Thus, analysis shifts from tracing individual paths to inferring the mechanisms that generate them, making causal inference essential. In this marketing context, extracting actionable insights remains challenging for both scholars and practitioners. When cause-and-effect relationships are identified, they can directly support more effective budget allocation decisions (Zhou et al., 2023).

Building on these challenges, our study offers primary contributions. First, we propose a theoretical approach that integrates Structural Equation Modelling (SEM) into a temporal causal discovery framework for simulating marketing attribution. Second, we introduce a causal-driven attribution framework that works entirely on aggregated impression-level data, solving the privacy and data-access limitations of user-level models. To the best of our knowledge, no existing approach fully recovers directional causal relationships and quantifies channel effects using only aggregated data.

The rest of the paper unfolds as follows. We begin with a review of the relevant literature, grounding our work in the methodological foundations of attribution modelling. We then outline our data simulation process and walk through the modelling approach in detail. This is followed by the presentation of key findings and an extensive model validation process. Finally, we close with a critical discussion of the results and what they mean for both future research and applied marketing practice.

[2]https://ads.tiktok.com/business/en/blog/beyond-click-attribution-views-impact-purchase
[3]https://forbusiness.snapchat.com/blog/evolving-measurement-neustar

4.2 Literature Review

4.2.1 Marketing Effectiveness and the Consumer Journey

Marketing effectiveness captures the extent to which marketing activities generate desired business outcomes such as sales, revenue growth, customer acquisition, or brand lift (Solcansky and Simberova, 2010; Raman et al., 2012). Although conceptually simple, evaluating effectiveness becomes increasingly complex once we account for how consumers behave and how marketers attempt to influence that behaviour (Chatterjee et al., 2025).

Classical theories framed the consumer journey as a linear progression from awareness to consideration to purchase (Singh and Yousuf, 2024), a structure that made sense in a world with limited media channels. However, these linear frameworks struggle to explain behaviour in contemporary ecosystems where consumers discover, evaluate, and engage with brands across search engines, social platforms, video environments, review sites, apps, and offline stimuli (Nam and Cho, 2025).

Digitalisation has expanded the number of consumer touchpoints and enabled multi-platform navigation. Omni-channel environments blur the boundaries between physical and digital experiences (Verhoef et al., 2015), while recent work shows that consumers routinely search on Google, watch YouTube videos, browse Instagram, and ultimately purchase offline (Santos and Gonçalves, 2024). Understanding this fragmented and recursive journey has therefore become central to enhancing marketing effectiveness (Kvíčala et al., 2024).

From a managerial perspective, resource allocation further complicates the challenge. Behavioural theories emphasise that managers operate under bounded rationality, seeking solutions that are "good enough" rather than fully optimal (Simon, 1955). Consequently, organisations make decisions under conditions of information overload, internal negotiation, and limited analytical capacity (March and Simon, 1993).

Cyert and March (2020) argue that firms do not optimise in the textbook sense. Instead, they rely on routines, satisficing, and sequential budget adjustments shaped by internal bargaining. As media channels proliferate and privacy constraints limit access to granular data, these behavioural constraints intensify. Recent work highlights the deep uncertainty marketing leaders face when evaluating investment effectiveness, thereby increasing demand for transparent and reliable measurement frameworks (Malter et al., 2020).

Due to behavioural complexity, fragmented data, and managerial uncertainty, attribution modelling has emerged as a central tool for accountability and optimisation.

4.2.2 Marketing Attribution

The Heuristic Attribution Era

The early days of marketing attribution relied on pre-defined rules that assigned credit to specific marketing touchpoints (Donndelinger and Ferguson, 2020). Google's attribution framework, developed between 2006 and 2015, introduced heuristic models such as first-interaction, linear, time-decay, position-based, and last-click (Gaur and Bharti, 2020). These rule-based models quickly became popular among practitioners and researchers in online marketing, and many of them remain in use today.

Among the various attribution models, last-click, which assigns all credit to the final touchpoint before conversion, gained significant traction as the main choice for both businesses and media companies. Despite well-documented conceptual weaknesses, it became the industry default. However, the adoption of last-click attribution increased even as criticism mounted (Kufile et al., 2023). This persistence reflects organisational behaviour: under uncertainty, firms gravitate toward simple and interpretable rules, even when these rules are inaccurate or questionable.

With the rise of digital media over the past decade, consumers have more options to engage with brands across multiple platforms. Likewise, brands are able to distribute their content and advertising message in more platforms (Dens and Poels, 2023). Therefore, heuristic models struggle to accurately capture channel contributions to key outcomes such as purchases (Tao et al., 2025).

First, it reflects reductive assumptions about consumer behaviour and fails to capture the full influence of earlier marketing interactions (El Mekkaoui and Benyoussef, 2024). Second, researchers highlight the increasingly non-linear nature of consumer journeys (Santos and Gonçalves, 2024). Contemporary consumer behaviour theory recognises highly fragmented journeys in which users move across multiple devices and platforms, interacting with both organic and paid content before making purchase decisions (Nam and Cho, 2025).

To mitigate these challenges, researchers came up with a different approach: data-driven models (Romero Leguina et al., 2020). For some, data-driven models are considered as "next generation attribution models" because they are grounded in mathematical and statistical methods rather than pre-defined rules.

The Data-Driven Attribution Era

As mentioned, consumers now engage with brands across multiple platforms, while brands have simultaneously expanded their presence across a growing number of digital touchpoints (Dens and Poels, 2023). This behavioural shift necessitated analysing entire paths rather than isolated events. Consequently, data-driven attribution models emerged as a methodological evolution, designed to process path data, meaning the full sequence of user touchpoints before conversion, and thereby capture the complexity of modern consumer journeys (Hui et al., 2009; Kannan et al., 2016). Several such methods now support marketing attribution decisions. Below, we highlight the approaches that have shaped the industry and discuss recent developments and emerging trends.

Shapley values, borrowed from cooperative game theory, provide an elegant mathematical framework for distributing credit across marketing touchpoints (Shapley et al., 1953). The method was first pioneered for marketing use in work that positioned it for practical application while acknowledging the limitations of observational data (Dalessandro et al., 2012). The approach is theoretically sound: it calculates each touchpoint's marginal contribution across all possible combinations and then assigns proportional credit. Later research enhanced computational efficiency, incorporated channel synergies, and extended the framework through hybrid modelling approaches (Yadagiri et al., 2015; Zhao et al., 2018; Kadyrov and Ignatov, 2019).

Markov chains offered an alternative framework by modelling consumer journeys as probabilistic state transitions (Pfeifer and Carraway, 2000). The "adgraph" concept was later introduced to represent journeys as graphs in which nodes capture ad events and edges encode transition probabilities (Archak et al., 2010). This approach effectively identified path dependencies and drop-off points, a conclusion supported by empirical validation comparing simple, forward, and backward Markov chains (Anderl et al., 2014). Subsequent extensions further demonstrated its applicability across industries (Kakalejčík et al., 2018; Sandeep et al., 2021*b*).

The integration of machine learning (ML) into attribution modelling provides a rigorous, data-driven approach capable of detecting patterns in customer behaviour and assessing the marginal influence of each touchpoint. Early work introduced bagged logistic regression and coined the term "data-driven attribution" (Shao and Li, 2011). Subsequent studies developed models distinguishing short- and long-term effects across cohorts (Breuer and Brettel, 2012), incorporating spillover effects through parametric and non-parametric approaches (Li and Kannan, 2014). Amazon's MTA approach leverages randomized controlled trials (RCTs) and ML models to allocate conversion credit across their ad touchpoints (Lewis et al., 2025). Other advances apply survival analysis to capture time-decay dynamics (Zhang et al., 2014) and utilise hazard-rate modelling to quantify the influence of individual ad exposures (Ji and Wang, 2017).

Building on ML-based advances, deep learning methods have recently been applied to the MTA problem, offering the ability to model nonlinear interactions, temporal dependencies, and user heterogeneity at scale. The

Deep Neural Net with Attention (DNAMTA) model integrates attention mechanisms to account for temporal effects and user characteristics, achieving strong predictive performance while producing interpretable touchpoint-level contribution estimates (Arava et al., 2018). The Dual-Attention Recurrent Neural Network (DARNN) further advances this idea by learning attribution weights directly from the conversion objective through a dual-attention architecture, dynamically allocating credit across a user's full exposure sequence (Ren et al., 2018). Furthermore, DeepMTA combines phased-LSTM networks with Shapley-based attribution to deliver accurate conversion prediction and transparent explanations of each touchpoint's marginal effect (Yang et al., 2020). Collectively, these approaches demonstrate how neural models capture richer temporal structure and provide more interpretable attribution estimates than traditional methods.

4.2.3 The Causal Imperative

Consider a marketer running a sales campaign across multiple channels, including email, social, and display. Although historical data may show higher conversion rates among customers exposed to more email messages, such correlations do not identify the true incremental effect of each touchpoint. Even if higher email volume coincides with higher conversions, it would be misleading to attribute all of those conversions to email. Many of these users may have already formed their purchase intent through other channels, such as search or social media, long before the email was delivered. Therefore, correlation-based attribution models can misallocate credit across channels because they are incapable of distinguishing true causal effects from correlations.

Causal inference has become increasingly central to marketing effectiveness research (Hair Jr. and Sarstedt, 2021), driven by advances in data availability, experimental design, and machine learning. This shift moves the field beyond correlational attribution toward frameworks that identify true cause–and–effect relationships, reflecting the growing recognition that understanding why marketing works is essential for credible measurement and strategic decision-making (Zhou et al., 2023). In this context, causal-based approaches to multitouch attribution estimate the incremental effect of each advertising touchpoint. Early work framed attribution as a causal estimation problem using Shapley values to allocate credit based on each ad's marginal effect. Kumar et al. (2020) introduce CAMTA, a causal recurrent neural network that accounts for time-varying confounders and reduces selection bias, although its use of click behaviour as pseudo-feedback may introduce bias. Yao et al. (2022) present CausalMTA, which mitigates confounding bias from static user attributes and dynamic behavioural features through counterfactual prediction. Tang (2024) advances this line of research with DCRMTA, a deep causal representation model that isolates causal features linking user behaviour to conversion while reducing confounding effects.

Different from the individual-level data used in most studies, Kireyev et al. (2016) utilise a dataset consisting only of aggregate consumer behaviour to examine the dynamic interaction between paid search and display advertising. They combine time-series analysis with Granger causality to identify cross-channel spillovers and broader advertising dynamics. However, Granger causality does not constitute true causal inference (Shojaie and Fox, 2022). It shows only that past values of one time series improve the prediction of another, not that one variable exerts a causal effect from the traditional causal-inference sense.

4.2.4 Toward a New Paradigm

Industry practice and academic literature increasingly converge on the same conclusion: today's environment exposes a critical measurement gap. Existing attribution frameworks break down when only aggregated or impression-level data are available, particularly for smaller or emerging channels that do not provide persistent user identifiers (Cui et al., 2021; Ghose and Todri-Adamopoulos, 2016). They also fail when user journeys are unobservable due to regulatory changes, when touchpoints do not generate clicks as in impression-first platforms, and when privacy constraints prohibit the collection of path-level behavioural data (as discussed in Section 4.1).

While a business could build complex API integrations across multiple platforms to capture user-level path data, it must carefully weigh the trade-off between adopting a new solution and maintaining the existing one at a high and potentially unsustainable infrastructure cost. Even with sufficient data, certain traditional MTA and deep learning models can be difficult to deploy due to their heavy computational requirements. Singh and Yousuf (2024) emphasise that computational efficiency is essential for modern customer-journey analytics, and models that require excessive computing resources are unlikely to be adopted in practice.

It is evident that there is a need for a new family of attribution methodologies. The shift from individual to aggregate data, from correlation to causation, and from clicks to impressions reflects more than a methodological evolution. It represents a fundamental reconceptualization of marketing measurement, one that aligns with how consumers behave and how privacy-conscious societies choose to govern their data. Our framework offers one step toward this reconceptualization.

Our work advances this direction by proposing a causal-driven attribution (CDA) methodology that (i) operates on aggregated, impression-level time-series data, (ii) is explicitly designed for privacy-restricted environments, (iii) accommodates channels that do not generate user-click signals, and (iv) identifies both direct and indirect causal relationships across marketing touchpoints.

4.3 Methodology

4.3.1 Data Generation

Data accessibility for advertisers has been significantly affected by regulatory and industry changes. Although such data are typically accessible via marketing APIs, effective use often requires substantial infrastructure available primarily to large firms (Simon, 2021). Additionally, Cui et al. (2021) clearly outlined the challenges associated with multi-touch attribution and how data limitations hinder its application.

Initially, we aimed to acquire a dataset suitable for comparing traditional multi-touch attribution methods with our proposed approach. However, over nearly two years of exploration, we were unable to find a suitable candidate. To our knowledge, existing studies in advertising and causal learning rely on only two publicly available datasets (Job et al., 2025), neither of which meets the requirements of this work. Due to the lack of ground truth in real-world causal marketing data, simulation has played a vital role in method development and evaluation.

Motivated by these challenges, we developed an approach that allowed us to generate a synthetic dataset. Our aim is not only to create a comprehensive and representative marketing dataset but moreover to be able to establish the so-called ground truth required for model validation, thereby bypassing the problem of insufficient data.

We acknowledge that real-world marketing data cannot be perfectly simulated due to hard-to-model factors, such as customer awareness and behaviour influenced by psychological processes. Accordingly, we introduce several simplifying assumptions to ensure that the proposed causal-driven attribution model remains applicable in practice. Under these assumptions, we incorporated baseline impressions, growth rate, and conversion rate for each channel, along with random noise to capture variability. Finally, we introduce cross-channel influence, allowing channels to act as confounders which is essential for causal inference conversation.

Causal DAG Graph

Before crunching any numbers, we require a high-level understanding of the causal relationships among marketing channels to define the data-generating scenario. Correlation describes statistical associations between variables but cannot answer questions like "Does changing the value of Channel A lead to a change in Channel B?" To simulate realistic marketing behaviour, we need a visual representation of the underlying causal structure, and this is where Directed Acyclic Graphs (DAGs) come into play.

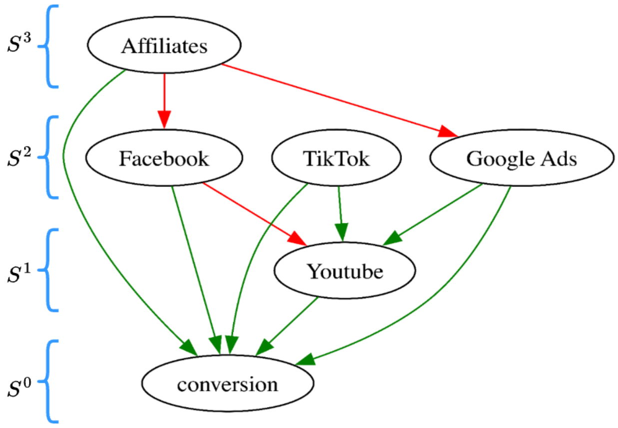


Figure 4.1: Example of a DAG's structure, including nodes: *Affiliates*, *Facebook*, *TikTok*, *Google Ads*, *YouTube*, and *conversion*. Edges represent directed causal influences: green arrows indicate positive values (putative amplification/halo), and red arrows indicate negative values (putative substitution/crowd-out). Nodes are grouped into causal layers as follows: S^3: *Affiliates*, S^2: *Facebook*, *TikTok*, *Google Ads*, S^1: *YouTube*, S^0: *conversion*.

Pearl (2009*b*) explains that DAGs provide a natural representation of causal structures. Causal effects arise from underlying, unobserved processes that can be viewed as equilibrium outcomes of behavioural equations forming a simplified model of the world. Graph notation captures this structure efficiently: nodes represent random variables generated by an underlying data-generating process, and directed edges encode causal influence. The acyclic property prevents feedback loops, thereby avoiding circular causation and preserving interpretability. Causal effects may be direct (e.g., $X \to Y$) or indirect (e.g., $X \to Z \to Y$), where influence is transmitted through a mediating variable.

A DAG $\mathcal{G}$ is a visual structure consisting of a set of nodes (vertices) $\mathbf{V}$ and a set of directed edges (arrows) $\mathcal{E}$. Each node $\mathbf{V}_i$ is associated with a random variable X_i. In a DAG, there are no directed paths that begin and end at the same node. Formally, no cycles exist such that a path starts at $\mathbf{V}_i$ and returns to $\mathbf{V}_i$ (either directly or indirectly). A finite set S denotes the causal layers (also known as a non-signalling set), which are subsets of nodes that are causally independent (i.e., nodes within the same layer cannot signal to one another) (Nashed et al., 2025). Fig. 4.1 illustrates the DAG structure, including its nodes, edges, and layered decomposition.

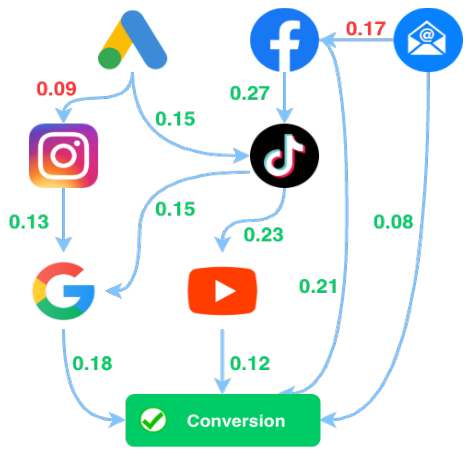


Figure 4.2: Example of a causal DAG showing the structural relationships among marketing channels. Some channels act as upstream drivers that influence conversions indirectly through intermediate touchpoints, while others have a direct impact on the outcome. All effects flow in a single direction with no circular feedback, providing a clear view of how influence moves through the system.

Recent literature highlights a growing use of DAGs as uncertainty-aware and interpretable tools for visualising causal structures (Xie et al., 2020). DAGs provide a transparent means of distinguishing causal influence from correlation. In marketing attribution, they map the structural relationships among channels and outcomes, clarifying how touchpoints interact, mediate, or confound one another along the consumer journey. Figure 4.2 illustrates an example DAG in a marketing context, illustrating how channels interact and ultimately influence conversions. Based on the generated DAGs, we subsequently construct the mathematical equations used for simulation.

Assumptions and Equations

We acknowledge that real marketing data cannot be perfectly simulated due to hard-to-model factors such as customer awareness and behaviour are influenced by psychological aspects. Additionally, since causal mechanisms in real-world marketing systems are known with high uncertainty (Runge et al., 2019), the simulation must reflect the properties of the real datasets on which the algorithms will be applied. This is necessary to ensure that our causal-driven attribution developed using simulated data can potentially translate to real-world scenarios.

A standard linear equation has been widely used to model the activity changes of each channel (Danaher and van Heerde, 2018; Zhao, Mahboobi and Bagheri, 2019). Suppose x_t represents the number of impressions for channel $X = (x_1, x_2...x_t, x_{t+1}...)$ at time t; the linear trend is defined as follows:

$$x_t = \alpha + \beta t + \epsilon_t, \tag{4.1}$$

where x_t is the channel value at time t, α represents the baseline impressions, β is the growth rate of the channel, and ϵ_t denotes the random noise.

The next step was to establish the cross-influence between channel coefficients. Taking two channels as an example, let X_1 represent *Search advertising* [4] and X_2 denote *Social advertising* [5]. Each channel still follows a linear model; however, they now influence each other directly within the same time period t (i.e., no lag), as follows:

$$\begin{aligned} x_{1,t} &= \alpha_1 + \beta_1 t + w_{12} x_{2,t} + \epsilon_{1,t}, \\ x_{2,t} &= \alpha_2 + \beta_2 t + w_{21} x_{1,t} + \epsilon_{2,t}, \end{aligned} \tag{4.2}$$

where $x_{1,t}$ and $x_{2,t}$ represent the impression amounts of *Social Ads* and *Search Ads* at time t, respectively; α_1 and α_2 denote the baseline impressions for each channel; β_1 and β_2 are their respective growth rates; $\epsilon_{1,t}$ and $\epsilon_{2,t}$ represent random noise terms.

More importantly, $\boldsymbol{w_{12}}$ measures how much *Social Ads* influence *Search Ads*, and $\boldsymbol{w_{21}}$ measures how much *Search Ads* affect *Social Ads*. These terms capture the directional causal effects between channels. Without them, the relationship would represent only correlation rather than causation.

Extending the idea to multiple channels, each channel's activity depends not only on its own trend but also on the concurrent activity of other chan-

[4] **Search advertising:** We refer to Search advertising (Search Ads) as a form of paid digital marketing in which advertisers bid to display sponsored messages alongside organic search results on search engines such as Google or Bing.

[5] **Social advertising:** We refer to Social advertising (Social Ads) as paid digital communication delivered through social networking platforms such as Facebook, Instagram, TikTok, or LinkedIn. It allows advertisers to target audiences based on demographic, psychographic, and behavioural data.

nels (without lag):

$$x_{i,t} = \alpha_i + \beta_i t + \sum_{j \neq i} w_{ij} x_{j,t} + \epsilon_{i,t}, \tag{4.3}$$

where α_i and β_i represent the baseline and growth trend of channel i, respectively, while w_{ij} measures how the current activity of channel j influences channel i. $\epsilon_{i,t}$ denotes the random noise term for channel i.

We then map impressions to conversions using the following relationship:

$$y_{t,i} = \gamma_0 + \sum_{i=1}^{N} \gamma_i x_{i,t} + \eta_t, \tag{4.4}$$

where $y_{t,i}$ denotes the number of conversions for channel i, γ_i represents the effectiveness of channel i, and η_t is the random noise term. This formulation captures the direct, concurrent (no-lag) causal influence among marketing channels.

Finally, the total number of conversions at each time step t is the sum of baseline (non-marketing) conversions and those attributed to all marketing channels.

$$C_t = B_t + \sum_{i=1}^{N} y_{t,i}, \tag{4.5}$$

where C_t represents the total conversions at time t, B_t represents baseline conversions not driven by advertising (e.g., organic or repeat behaviour), and $y_{t,i}$ denotes conversions generated by marketing channel i at the same time step.

This simulation integrates organic and paid contributions to overall performance, allowing the model to disentangle the inherent baseline trend from the incremental impact of each channel. It provides a comprehensive view of how marketing efforts collectively contribute to total conversions while maintaining interpretability at the channel level.

Figure 4.3 illustrates the proposed simulation framework for generating synthetic causal marketing attribution data. The process begins with user-defined inputs specifying the simulation context, including (1) the set of marketing channels and their activities (e.g., impressions) and the tar-

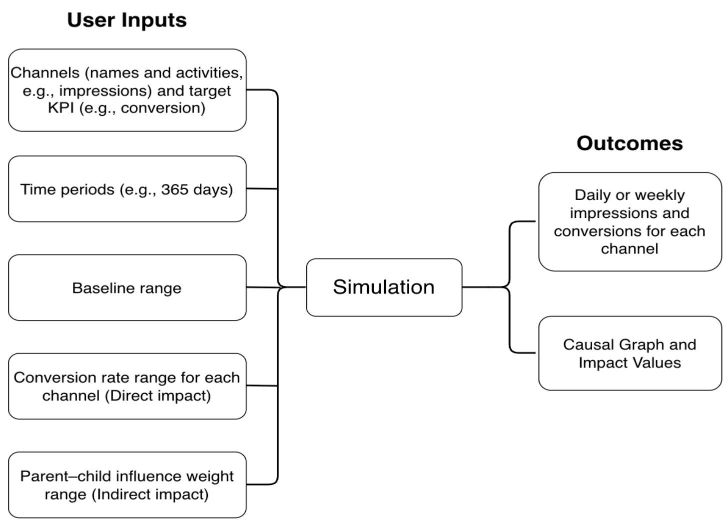


Figure 4.3: Overview of the simulation framework.

get key performance indicator (KPI), such as conversions; (2) the simulation horizon (e.g., 365 days); (3) a baseline range capturing conversions independent of media activity; (4) channel-specific conversion-rate ranges governing the direct impact of media exposure; and (5) parent–child influence weight ranges that capture indirect effects arising from causal dependencies among channels.

Given these inputs, the simulation module produces two categories of outputs. The first is the time-series data of daily or weekly impressions and conversions for each channel (as shown in Fig. 4.4), reflecting both direct and mediated effects. The second is a causal graph with corresponding impact values that quantify the impact values and direction of relationships among channels and between channels and the target KPI (Fig. 4.5). Together, these outputs enable the analysis of causal influence and attribution in a controlled, reproducible environment.

4.3.2 Modelling

Attribution is fundamentally a cause-and-effect problem. Understanding causality in attribution is central to decision-making under uncertainty (Kelly et al., 2018). A causal model identifies and maps the qualitative cause-and-effect relationships and combines them with quantitative information about the strengths of those relationships. Therefore, our approach

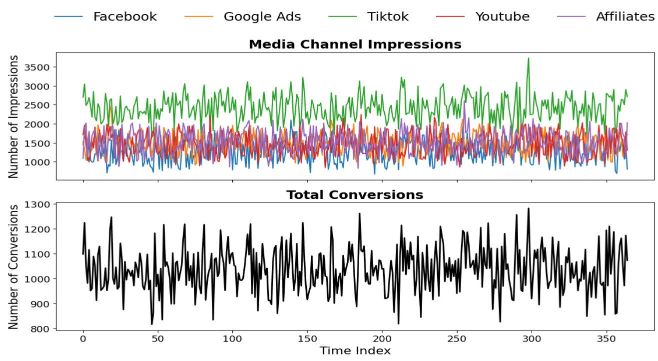


Figure 4.4: A random simulated data illustrating marketing channel impressions and conversions within 365 days. The top panel presents the impression trends for five marketing channels, including Google Ads, Affiliates, TikTok, Facebook, and YouTube. Each series follows a distinct stochastic pattern generated from a linear growth process with random noise and inter-channel influence.

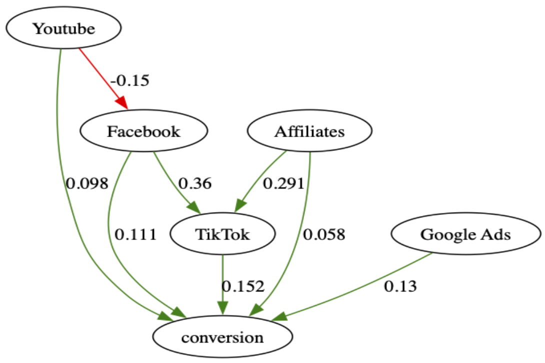


Figure 4.5: Example DAG output illustrating multi-channel interactions from simulated data. Nodes represent marketing channels (Affiliates, Google Ads, TikTok, Facebook, YouTube) and the target outcome (conversions). Directed arrows denote causal influence from cause to effect, with green and red arrows indicating positive and negative causal weights, respectively.

uses a two-stage framework: first, we discover the causal relationships among marketing channels; then we estimate the causal influence (weights) of each channel on the target KPI (e.g., conversions).

Causal DAG Discovery

Through simulation, causal relationships between channels and conversions can be observed directly. In reality, however, businesses typically have access only to time-series data on channel activity and aggregate conversions, similar to the simulated multi-channel dataset shown in Table 4.1. Incrementality approaches are commonly used to infer causality, as it identifies what actually caused a lift in performance. A causal DAG can be constructed using prior xpert knowledge and experimental tests such as lift studies or A/B experiments, which are costly and time-consuming. Consequently, discovering causal structure directly from observational data remains a challenging yet valuable problem. In this work, we apply PCMCI, an automated procedure for discovering causal graphs based on graphical causal modeling and conditional independence principles. Moreover, PCMCI scales effectively to large time series datasets and accommodates linear, nonlinear, and time-delayed dependencies.

Table 4.1: Schema of the simulated multi-channel marketing dataset.

Column	Type	Description
`index`	integer	Time step index (e.g., day or week).
`conversion`	float	Total conversions across all marketing channels.
`channel_1_impression`	float	Impressions from marketing channel 1.
`channel_2_impression`	float	Impressions from marketing channel 2.
...	...	...
`channel_n_impression`	float	Impressions from marketing channel n.

Causal Network Discovery Using PCMCI Consider an underlying multivariate time series $\mathbf{X}_t = \left(X_t^{(1)}, X_t^{(2)}, \ldots, X_t^{(d)}\right)$, where each component represents the daily activity of a marketing channel. Our objective is to infer the causal structure governing how past channel activities influence current outcomes. Time lag τ, also know as "time-delay causality", refers to the temporal interval between a cause and its effect. We assume finite-order temporal dependencies up to a maximum lag $\tau_{\max}$ and seek, for each variable $X_t^{(i)}$, to estimate the set of causal parents

$$\mathcal{P}(X_t^{(i)}) \subseteq \left\{X_{t-\tau}^{(j)} \;\middle|\; 1 \leq j \leq d,\ 1 \leq \tau \leq \tau_{\max}\right\}. \tag{4.6}$$

PCMCI proceeds in two stages. In the first stage, a PC_1 condition-selection step identifies relevant conditioning sets $\widehat{\mathcal{P}}(X_t^j)$ for each variable X_t^j. Starting from all lagged variables up to $\tau_{\max}$, variables that are conditionally independent of the target given subsets of the remaining candidates are removed using a conditional independence test (e.g., partial correlation, ParCorr) performed at significance level α_{PC}.

In the second stage, PCMCI applies the Momentary Conditional Independence (MCI) test to evaluate whether a lagged variable $X^i_{t-\tau}$ is a causal parent of X^j_t. For each variable pair, MCI tests

$$X^i_{t-\tau} \not\perp\!\!\!\perp X^j_t \mid \widehat{\mathcal{P}}(X^j_t) \setminus \{X^i_{t-\tau}\},\ \widehat{\mathcal{P}}(X^i_{t-\tau}), \tag{4.7}$$

using an appropriate conditional independence test (e.g., ParCorr) at significance level α_{MCI}. If the null hypothesis of conditional independence is rejected, the directed lagged edge $X^i_{t-\tau} \to X^j_t$ is included in the estimated causal graph. Pseudocode 4.3.2 summarizes the full PCMCI procedure implemented in this study. PCMCI for multivariate time series causal discovery [1]

Input: Multivariate time series $\{\mathbf{X}_t\}_{t=1}^T$ with d variables; maximum lag $\tau_{\max}$; thresholds α_{PC} and α_{MCI}. **Output:** Set of directed lagged edges $\mathcal{E}$.

Construct candidate parent sets:

$$\mathcal{P}_0(X_t^{(i)}) = \left\{X_{t-\tau}^{(j)} \,\middle|\, 1 \le j \le d,\ 1 \le \tau \le \tau_{\max}\right\}.$$

Stage 1: PC condition selection

$i = 1$ to d $\mathcal{P}(X_t^{(i)}) \leftarrow \mathcal{P}_0(X_t^{(i)})$ $k \leftarrow 0$ conditioning sets of size k exist $X_{t-\tau}^{(j)} \in \mathcal{P}(X_t^{(i)})$ For all $\mathbf{Z} \subseteq \mathcal{P}(X_t^{(i)}) \setminus \{X_{t-\tau}^{(j)}\}$ with $|\mathbf{Z}| = k$, test:

$$H_0 :\ X_{t-\tau}^{(j)} \perp\!\!\!\perp X_t^{(i)} \mid \mathbf{Z}.$$

any $p > \alpha_{\text{PC}}$ Remove $X_{t-\tau}^{(j)}$ from $\mathcal{P}(X_t^{(i)})$ $k \leftarrow k+1$

Stage 2: MCI testing

$\mathcal{E} \leftarrow \emptyset$

$i = 1$ to d $X_{t-\tau}^{(j)} \in \mathcal{P}(X_t^{(i)})$ Define the MCI conditioning set:

$$\mathbf{Z}_{\text{MCI}} = \mathcal{P}(X_t^{(i)}) \setminus \{X_{t-\tau}^{(j)}\}.$$

Test the dependence condition:

$$X_{t-\tau}^{(j)} \not\perp\!\!\!\perp X_t^{(i)} \mid \mathbf{Z}_{\mathrm{MCI}}.$$

$p \leq \alpha_{\mathrm{MCI}}$ Add edge $(X_{t-\tau}^{(j)} \to X_t^{(i)})$ to $\mathcal{E}$

$\mathcal{E}$

Causal Effect Estimation

In this section, we demonstrate how our proposed method quantifies causal effects under linear structural dependencies. To estimate the causal contribution of each marketing channel to conversions, we use a Structural Causal Model implemented with the `DoWhy` framework [6]. An SCM represents each variable X_i through a structural equation

$$X_i = f_i\big(\mathcal{P}(X_i^t), N_i\big), \tag{4.8}$$

where $\mathcal{P}(X_i^t)$ denotes the causal parents of X_i^t and N_i is an exogenous noise term. Once the DAG is inferred, each causal mechanism f_i is estimated from the data.

In causal inference, several estimands characterize the effect of an intervention, including the Average Treatment Effect (ATE), the Conditional Average Treatment Effect (CATE), the Individual Treatment Effect (ITE), the Average Treatment Effect on the Treated (ATT), the Average Treatment Effect on the Controls (ATC), and the Local Average Treatment Effect (LATE) (Imbens and Rubin, 2015; Cheng et al., 2022). Each estimand contrasts potential outcomes under treatment ($T = 1$) and control ($T = 0$), differing in how effects are averaged and which subpopulations are targeted.

In our setting, each observation is generated by a SCM and corresponds to a specific configuration of channel impressions and contextual covariates. Therefore, the causal effect of a marketing channel on conversions must be evaluated *conditional* on these observed features. The CATE is well suited for this purpose, as it measures how the expected conversion

[6] `https://www.pywhy.org/dowhy`

outcome changes when a specific channel is intervened upon (e.g., setting its impression level to zero) while holding all other covariates fixed.

To quantify the effect of removing a marketing channel, we formalize the intervention using the do-operator. Let X_j denote the activity of channel j, Y_t the conversion outcome, and Z_t the set of all other observed covariates at time t. We consider a binary intervention in which the channel is either kept at its normalized active level ($X_j = 1$) or removed ($X_j = 0$). The CATE of channel j at covariate profile z is then

$$\Delta_j(z) = E[Y_t \mid \mathrm{do}(X_j = 1),\, Z_t = z] - E[Y_t \mid \mathrm{do}(X_j = 0),\, Z_t = z]\,, \tag{4.9}$$

and we approximate these conditional expectations by generating interventional samples from the fitted SCM while conditioning on the empirical distribution of Z_t.

Because these counterfactual expectations are not available in closed form, we estimate them using interventional Monte Carlo sampling. For each run i, we generate two sets of interventional samples conditioned on the observed covariates $Z_t = z$: one under the intervention $\mathrm{do}(X_j = 1)$ and one under $\mathrm{do}(X_j = 0)$. This yields empirical means $\overline{Y}_{\mathrm{do}(X_j=1),\,i}$ and $\overline{Y}_{\mathrm{do}(X_j=0),\,i}$. Over n runs, the conditional causal effect of channel j at covariate profile z is then estimated by

$$\widehat{\Delta}_j(z) = \frac{1}{n}\sum_{i=1}^{n}\left(\overline{Y}_{\mathrm{do}(X_j=1),\,i} - \overline{Y}_{\mathrm{do}(X_j=0),\,i}\right), \tag{4.10}$$

which converges to the true CATE $\Delta_j(z)$ as n increases. Repeating this sampling procedure over n runs yields a stable estimate of the conditional causal effect. This Monte Carlo procedure also enables uncertainty quantification without requiring parametric assumptions on the structural equations.

To express the causal effect on a comparable scale across channels, we compute a per-unit causal contribution by normalizing the estimated effect by the average activity level of channel j:

$$\text{ACE per unit}(j) = \frac{\widehat{\Delta}_j(z)}{E[X_j]}. \tag{4.11}$$

This quantity represents the incremental causal contribution of a single unit of channel activity (e.g., one impression) to expected conversions.

4.3.3 Evaluation Metrics

Causal DAG Discovery

To quantify the DAG structures learned by our proposed causal-discovery algorithm, we adopt two complementary families of metrics: (i) *edge-based graph classification measures* to compare the number of edges that differ between two graphs (Cheng et al., 2022), and (ii) *graph-distance-based measures*, which evaluate how well the learned graph supports valid identifying formulas for effects, and. Below we go through each measure in detail.

The edge-based graph distance will be based on the intuition that directional adjacency relations can be treated as a binary classification problem. Therefore, a variety of metrics in classification tasks can be used such as AUC, True Positive Rate (TPR), False Positive Rate (FPR), and F_β.

True Positive Rate and False Positive Rate We consider a finite set of random variables $X_1, \ldots, X_D$ with index set $\mathbf{V} = \{1, \ldots, D\}$. A graph $\mathcal{G} = (\mathbf{V}, \mathcal{E})$ consists of the node set $\mathbf{V}$ and an edge set $\mathcal{E} \subseteq \mathbf{V} \times \mathbf{V}$.

Let $\mathcal{G}_{\text{true}} = (\mathbf{V}, \mathcal{E}_{\text{true}})$ denote the ground truth DAG, and let $\mathcal{G}_{\text{pred}} = (\mathbf{V}, \mathcal{E}_{\text{pred}})$ denote the predicted DAG.

The true positive rate (TPR) is defined as the ratio of correctly recovered edges, meaning the edges that appear in both $\mathcal{E}_{\text{pred}}$ and $\mathcal{E}_{\text{true}}$, relative to the total number of edges in the ground truth graph. Formally,

$$\text{TPR} = \frac{|\mathcal{E}_{\text{pred}} \cap \mathcal{E}_{\text{true}}|}{|\mathcal{E}_{\text{true}}|}. \tag{4.12}$$

The false positive rate (FPR) measures the proportion of incorrectly added edges, meaning the edges that appear in $\mathcal{E}_{\text{pred}}$ but not in $\mathcal{E}_{\text{true}}$, relative to the number of edges that are not present in the ground truth graph.

Formally,

$$\mathrm{FPR} = \frac{|\mathcal{E}_{\text{pred}} \setminus \mathcal{E}_{\text{true}}|}{|(\mathbf{V} \times \mathbf{V}) \setminus \mathcal{E}_{\text{true}}|}, \tag{4.13}$$

where $\mathbf{V} \times \mathbf{V}$ denotes the set of all possible directed edges among the variables, excluding self loops if desired.

TPR and FPR are computed from the p-value matrices after applying a p% (e.g., 5%) significance threshold. TPR measures the fraction of true causal links that the method correctly identifies, while FPR measures the fraction of absent links that the method incorrectly labels as causal. Formally, TPR is the number of correctly detected links divided by the number of true links, and FPR is the number of falsely detected links divided by the number of absent links. Ideally, a causal discovery algorithm should recover true causal links with high sensitivity while keeping false positive (incorrect link) detections under control.

Area Under the ROC Curve Area under ROC curve (AUC) is the area under the curve of recall versus FPR at different thresholds. The AUC quantifies the ability of the model to distinguish true causal links from non-links based on the score matrices returned by PCMCI. Each score matrix contains non-negative values, where a larger entry at position (i, j) indicates greater confidence in a causal edge $i \to j$. The corresponding ground-truth label is binary (1 for a true causal link, 0 otherwise). Higher AUC values, with a maximum of 1, indicate stronger separation between true and false links.

F-Measure (F_β) The F-measure, also computed on thresholded p-value matrices, evaluates the harmonic mean of precision and recall. It ranges from 0 (poor performance) to 1 (perfect precision and recall). In this study, we use $F_{0.5}$, which places greater emphasis on precision than recall, reducing the impact of false negatives. This makes $F_{0.5}$ particularly suitable when the cost of false positives is higher or when identifying spurious edges is more detrimental than missing weak ones.

Nevertheless, the number of differing edges between two graphs does not capture how the graphs differ with respect to the identifying formulas they imply for causal effects. Therefore, edge-based metrics cannot quantify whether a predicted graph preserves the causal semantics encoded by the true graph. In particular, two graphs may differ in several edges yet induce the same interventional distributions, while another pair may differ by only a few edges but imply fundamentally different causal effects. To evaluate causal fidelity rather than purely structural similarity, we additionally employ graph-distance-based measures that compare graphs in terms of the causal relationships they encode, as described in the next section.

Structural Hamming Distance The Structural Hamming Distance (SHD) (Cheng et al., 2022) is one of the most commonly used metrics for evaluating the accuracy of a learned causal graph. Intuitively, it counts the number of edge insertions, deletions, and orientation reversals required to transform the learned graph into the true graph. Lower SHD values indicate more accurate causal reconstruction, whereas higher values imply poorer causal discovery performance.

Given a true causal graph G_{true} and a predicted (learned) graph G_{pred}, both defined over the same vertex set $\mathbf{V}$, we define their adjacency matrices $A_{\text{true}}, A_{\text{pred}} \in \{0,1\}^{D\times D}$ such that $A[i,j] = 1$ if there is an edge $X_i \to X_j$.

The Hamming distance between binary vectors measures the number of directed edges that differ between the two adjacency matrices. Formally,

$$\begin{aligned} \mathrm{SHD}(G_{\text{true}}, G_{\text{pred}}) &= \sum_{i=1}^{D}\sum_{j=1}^{D} |A_{\text{true}}[i,j] - A_{\text{pred}}[i,j]| \\ &= \|A_{\text{true}} - A_{\text{pred}}\|_1. \end{aligned} \tag{4.14}$$

Alternatively, for each unordered pair of nodes (i,j) with $i<j$, define

$$\delta_{ij} = \begin{cases} 1, & \text{if } A_{\text{true}}[i,j] \neq A_{\text{pred}}[i,j] \\ & \text{or } A_{\text{true}}[j,i] \neq A_{\text{pred}}[j,i], \\ 0, & \text{otherwise.} \end{cases} \tag{4.15}$$

Then,

$$\mathrm{SHD}(G_{\text{true}}, G_{\text{pred}}) = \sum_{1 \le i < j \le D} \delta_{ij}. \tag{4.16}$$

The normalized variant scales the distance by the number of possible unordered node pairs:

$$nSHD(G_{\text{true}}, G_{\text{pred}}) = \frac{1}{\binom{D}{2}} \sum_{1 \le i < j \le D} \delta_{ij}, \tag{4.17}$$

so that $n\mathrm{SHD} \in [0, 1]$. This normalization expresses the proportion of differing edges relative to the total possible pairwise relations.

Although the SHD implementation is simple and efficient to compute, it only reflects local structural differences and ignores causal equivalence or d-separation patterns. Even a very small structural change (e.g., flipping a single edge $A \to B \to C$ to $A \leftarrow B \to C$) can completely alter a graph's causal interpretation even when SHD suggests that the two graphs are nearly identical. Consequently, SHD has been criticized for failing to reflect the full causal implications of structural differences between graphs. This limitation has motivated the development of more causally informed metrics such as the Structural Intervention Distance (SID), discussed below.

Structural Intervention Distance The Structural Intervention Distance (SID) counts the number of ordered node pairs (i, j) for which the interventional distributions implied by the true DAG and the learned DAG differ. To understand SID, we introduce the intervention distribution.

Given a directed acyclic graph (DAG) $\mathcal{G}$ over variables $\mathbf{X} = (X_1, \dots, X_p)$ with joint distribution $\mathcal{L}(\mathbf{X})$, the parents of a node X_j are defined as

$$\mathrm{PA}_j^{\mathcal{G}} = \{\, X_i : (i, j) \in \mathcal{E} \,\}, \tag{4.18}$$

meaning that X_i is a parent of X_j whenever the edge $X_i \to X_j$ is present.

Pearl's do-operator formalises interventions by replacing the conditional distribution of the intervened node. Intervening on X_j with distribution

$\tilde{p}(x_j)$ yields

$$p(x_1, \ldots, x_p \mid do(X_j = \tilde{p}(x_j))) = \prod_{i \neq j} p(x_i \mid x_{\mathrm{PA}_i}) \cdot \tilde{p}(x_j), \tag{4.19}$$

whenever $p(x_1, \ldots, x_p) > 0$, and zero otherwise.

A point intervention $do(X_j = \tilde{x}_j)$ is obtained by choosing $\tilde{p}(x_j) = \delta_{x_j, \tilde{x}_j}$:

$$p(x_1, \ldots, x_p \mid do(X_j = \tilde{x}_j)) = \prod_{i \neq j} p(x_i \mid x_{\mathrm{PA}_i}) \cdot \delta_{x_j, \tilde{x}_j}. \tag{4.20}$$

Often we are interested only in the marginal effect of X_j on another variable X_i. If X_i is not a parent of X_j, the parent adjustment formula (Theorem 3.2.2 in (Pearl, 2009*b*)) gives:

$$p(x_i \mid do(X_j = \tilde{x}_j)) = \sum_{x_{\mathrm{PA}_j}} p(x_i \mid \tilde{x}_j, x_{\mathrm{PA}_i})\, p(x_{\mathrm{PA}_j}). \tag{4.21}$$

SID compares a true graph $\mathcal{G}_{\text{true}}$ with an estimated graph $\mathcal{G}_{\text{pred}}$ by counting the number of node pairs (i, j) for which the interventional distributions differ:

$$\begin{aligned} \mathrm{SID}(\mathcal{G}_{\text{true}}, \mathcal{G}_{\text{pred}}) = \Big| \big\{ (i, j) : & \, p_{\mathcal{G}_{\text{pred}}}(X_j \mid do(X_i)) \\ & \neq \, p_{\mathcal{G}_{\text{true}}}(X_j \mid do(X_i)) \big\} \Big|. \end{aligned} \tag{4.22}$$

A pair (i, j) is called *good* if the intervention distribution obtained from the learned graph matches that from the true graph for all observational distributions $\mathcal{L}(\mathbf{X})$. Otherwise, the pair contributes one unit to the SID score. In this way, SID evaluates errors in causal ordering, whereas SHD counts only incorrect edges.

For a graph with V nodes, the number of ordered node pairs (i, j) with $i \neq j$ is

$$N_{\text{pairs}} = V(V - 1). \tag{4.23}$$

The normalized Structural Intervention Distance is defined as

$$\mathrm{nSID}(\mathcal{G}_{\text{true}}, \mathcal{G}_{\text{pred}}) = \frac{\mathrm{SID}(\mathcal{G}_{\text{true}}, \mathcal{G}_{\text{pred}})}{V(V - 1)}. \tag{4.24}$$

Thus, nSID reflects the proportion of incorrectly estimated causal effects, with values near 0 indicating high accuracy and values near 1 indicating poor performance.

Causal Effect Estimation

After estimating the CATE for each marketing channel (as shown in Section 4.3.2), we evaluate the accuracy of the causal-driven attribution model using three complementary metrics.

Relative Root Mean Square Error The relative root mean square error (RRMSE) measures the scale-normalized deviation between the estimated causal effects $\hat{\tau}_i$ and the ground-truth effects τ_i:

$$\text{RRMSE} = \frac{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{\tau}_i - \tau_i)^2}}{\frac{1}{n}\sum_{i=1}^{n}|\tau_i|}, \tag{4.25}$$

where n is the number of channels. This metric captures overall estimation error relative to the magnitude of true effects.

The interpretation of the RRMSE depends entirely on the context, domain, and units of the outcome variable. Different research areas adopt different thresholds for what constitutes an acceptable level of normalized error. For example, in the energy and agricultural modeling literature, model accuracy is often considered *excellent* when RRMSE $< 10\%$, *good* when $10\% \leq$ RRMSE $< 20\%$, *fair* for $20\% \leq$ RRMSE $< 30\%$, and *poor* when RRMSE $\geq 30\%$ (Jamieson et al., 1991; Despotovic et al., 2016). These thresholds illustrate that RRMSE benchmarks are not universal but rather depend on the application context and the expected signal-to-noise ratio in the data.

Mean Absolute Percentage Error The mean absolute percentage error (MAPE) expresses the average proportional deviation:

$$\text{MAPE} = \frac{100\%}{n}\sum_{i=1}^{n}\left|\frac{\hat{\tau}_i - \tau_i}{\tau_i}\right|. \tag{4.26}$$

RRMSE and MAPE capture complementary aspects of estimation accuracy. RRMSE gives more weight to larger errors because it is based on *squared deviations* that are normalized by the scale of the true effects, making it sensitive to large misestimations. In contrast, MAPE provides an interpretable percentage-based measure of accuracy by averaging *absolute proportional errors*. For instance, a MAPE of 5% indicates that the estimated causal effects deviate from the true effects by only 5% on average.

Spearman Rank Correlation To evaluate whether the attribution model preserves the relative importance of marketing channels, we compute the Spearman rank correlation between the ground-truth causal effects and the model's estimated causal effects (Cornell et al., 2024).

Let C denote the list of marketing channels sorted by their true causal effect, and C_p the corresponding ranking produced by the attribution model. For each channel x, we define its rank in the true and predicted lists as

$$\begin{aligned} r_C(x) &= \text{position of channel } x \text{ in } C, \\ r_{C_p}(x) &= \text{position of channel } x \text{ in } C_p. \end{aligned} \tag{4.27}$$

From these ranks, we construct two ranked vectors:

$$\begin{aligned} R_C &= \big(r_C(x_1),\, r_C(x_2),\, \dots,\, r_C(x_n)\big), \\ R_{C_p} &= \big(r_{C_p}(x_1),\, r_{C_p}(x_2),\, \dots,\, r_{C_p}(x_n)\big), \end{aligned} \tag{4.28}$$

where n is the total number of marketing channels.

The Spearman rank correlation coefficient ρ between the two ranked lists is defined as

$$\rho = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)}, \tag{4.29}$$

where $d_i = r_C(x_i) - r_{C_p}(x_i)$ is the difference in rank for the i^{th} channel, and n is the total number of marketing channels.

Spearman's ρ takes values between -1 and 1. A value of $\rho = 1$ indicates perfect agreement, meaning the model recovers the exact ordering of true causal contributions across marketing channels. A value of $\rho = -1$ means the estimated ranking is the exact reverse of the true ranking. A value of $\rho = 0$ indicates no monotonic relationship between the estimated and true rankings. In our study, a high Spearman correlation signifies that the model correctly identifies which channels are more or less influential, even if the estimated magnitudes differ. This property is particularly important for resource allocation and budget optimization, where the correct ordering of channel importance often matters more than precise effect magnitudes.

4.4 Results

This study uses *Tigramite*[7], an open-source Python package for reconstructing graphical models (conditional independence graphs) from multivariate time series based on the PCMCI framework, and *DoWhy*[8] for estimating causal effects.

The algorithm proceeds in two steps. First, a PC step screens candidate parent sets using a permissive threshold ($\alpha_{\text{PC}} = 0.2$) to favour recall and preserve potentially relevant dependencies. Second, a Momentary Conditional Independence (MCI) test re-evaluates each retained link at a stricter significance level ($\alpha_{\text{MCI}} = 0.05$), ensuring robustness and suppressing spurious associations. This two-step procedure balances false negatives and false positives, yielding stable edge sets across repeated runs.

To ensure a DAG causal structure, bidirectional links are pruned by retaining the edge whose time-lagged effect has the greater magnitude, and any remaining cycles are resolved using a priority rule that preserves edges directed into the main outcome variable (e.g., conversions). This enforces causal directionality toward the dependent variable of interest.

4.4.1 Optimal Time-delay Causality

Nevertheless, a major challenge when using PCMCI is specifying the appropriate time lag needed to reliably estimate causal links. We do not know in advance which time lag best reflects the underlying causal relationships in

[7] https://jakobrunge.github.io/tigramite/
[8] https://github.com/py-why/dowhy

the simulated data. Therefore, an empirical assessment is necessary to identify which lag values allow PCMCI to most reliably recover the underlying causal structure in our simulated environment.

To this end, we explore temporal lags up to $\tau_{\max} = 60$ and evaluate model performance across 1,000 simulated datasets. For each lag value τ, we compute predictive and structural metrics, including AUC, TPR, FPR, F0.5, SHD, and SID, and summarize their empirical distributions using kernel density estimates (KDEs), shown in Fig. 4.6. Each ridge represents 1,000 simulations at a fixed lag and illustrates how both the central tendency and variability of each metric evolve as τ increases.

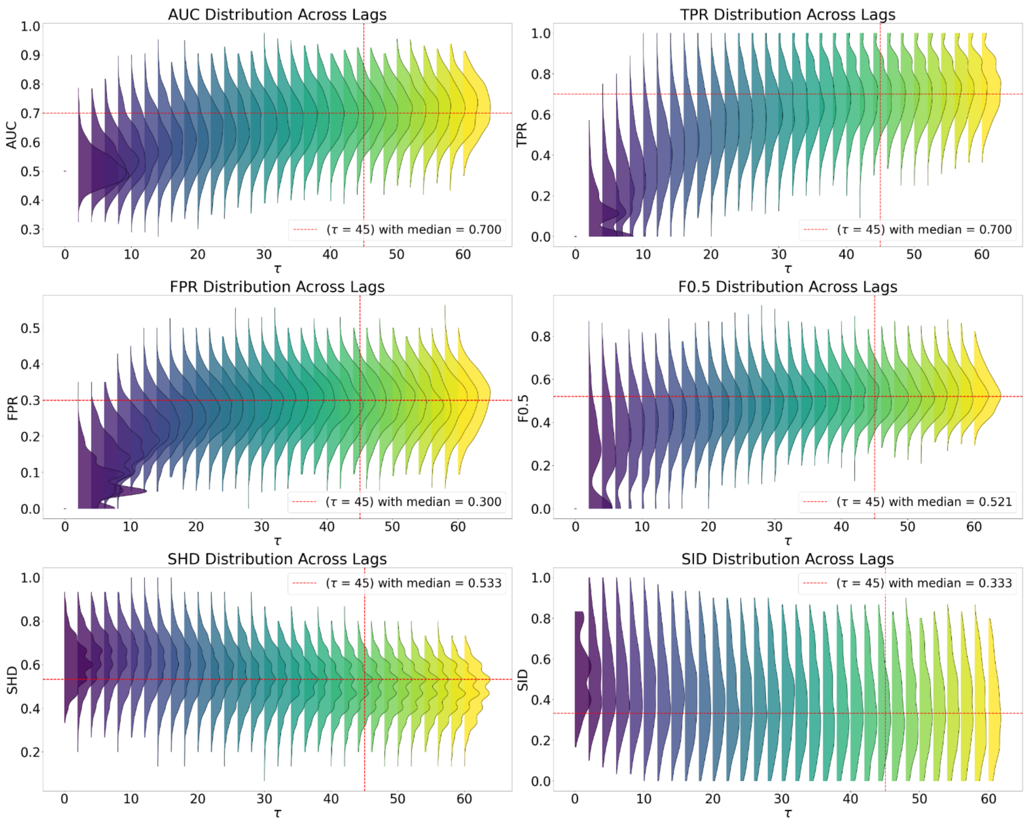


Figure 4.6: Kernel density estimates (KDEs) of six evaluation metrics, including AUC, TPR, FPR, F0.5, SHD, and SID across lag values (τ). Each subplot shows how the empirical distribution of the metric varies as τ increases from 0 to 60, using 1000 simulation seeds for each τ. The viridis color gradient is used to visually distinguish KDE ridges at different lag values; color does not encode density or any statistical quantity. Instead, the horizontal width of each ridge reflects the density and spread of metric values at a given lag. The red dashed vertical line marks $\tau = 45$, and the red dashed horizontal line denotes the corresponding median metric value at that lag.

Across all six metrics, a consistent pattern emerges: performance steadily improves with increasing lag, reaches its optimum around $\tau = 45$, and then either plateaus or slightly deteriorates. At $\tau = 45$, the KDE ridges indicate three desirable properties: *(i)* the best concentration of high-performing metric values, *(ii)* the lowest variability across seeds, and *(iii)* a balanced trade-off between predictive accuracy (AUC, TPR, F0.5) and structural recovery (SHD, SID). In addition, the FPR remains well controlled around this lag, meaning it does not increase substantially despite gains in sensitivity (TPR). Such behaviour is consistent with a saturation effect, where additional lag increases no longer introduce new false positives because most meaningful dependencies have already been captured. This confluence of high performance and low uncertainty identifies $\tau = 45$ as the optimal lag value τ_{max}.

In conclusion, this result indicates that the simulated system exhibits a characteristic time delay of approximately 45 time steps between channel impressions and observed conversions. At this lag, PCMCI most effectively captures the temporal propagation of causal influence across variables. For this reason, all subsequent causal attribution experiments adopt $\tau_{max} = 45$ as the default time-delay configuration.

4.4.2 Causal-Driven Attribution Analysis

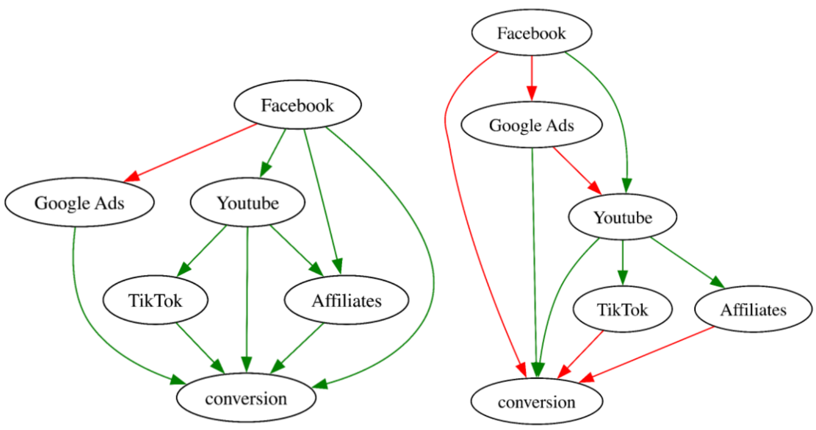
Illustrative Example

Causal DAG Discovery Fig. 4.7 compares the ground-truth causal graph (left) with the graph predicted by the PCMCI algorithm (right). The true DAG is generated from our simulated structural model (with *seed* value 155 in this illustration), where each directed edge corresponds to a known causal relationship and all effects on the target variable, *conversion*. Of course, the choice of seed is arbitrary, and any random seed would produce a valid instance of the simulated structural model. Overall, our estimator successfully recovers the major components of the underlying structure.

According to graph-distance–based metrics, the comparison determined an SHD of 0.133 and an SID of 0.10, indicating strong agreement with the ground-truth DAG in both its structural configuration and its implied interventional behaviour. These low values suggest that only a small number of edge discrepancies remain and that the learned graph preserves most of the true causal dependencies.

From a classification perspective, we also obtained a strong performance. The TPR is 0.90, which means that 90% of actual causal edges were correctly identified. The FPR is relatively low at 0.05, indicating that the predicted graph is well calibrated in its edge decisions. Furthermore, the AUC is 0.925, showing that our method achieves a high level of discriminative power when ranking edges by their likelihood of being causal. The $F_{0.5}$ score is 0.90, which confirms that the majority of the identified edges are accurate and not attributable to false detections.

Although our causal graph discovery approach recovers much of the true structure, it also introduces minor errors. For example, it fails to detect a true edge from *Facebook* $\rightarrow$ *Affiliates*, and it incorrectly introduces spurious edges such as *Google Ads* $\rightarrow$ *YouTube*. Additionally, the learned graph contains certain sign errors, where the direction of causal influence is correct but the estimated effect is negative rather than positive. To address these issues, a post-estimation refinement step will be applied to improve the accuracy of the predicted DAG.



(a) True Causal DAG (b) Predicted Causal DAG

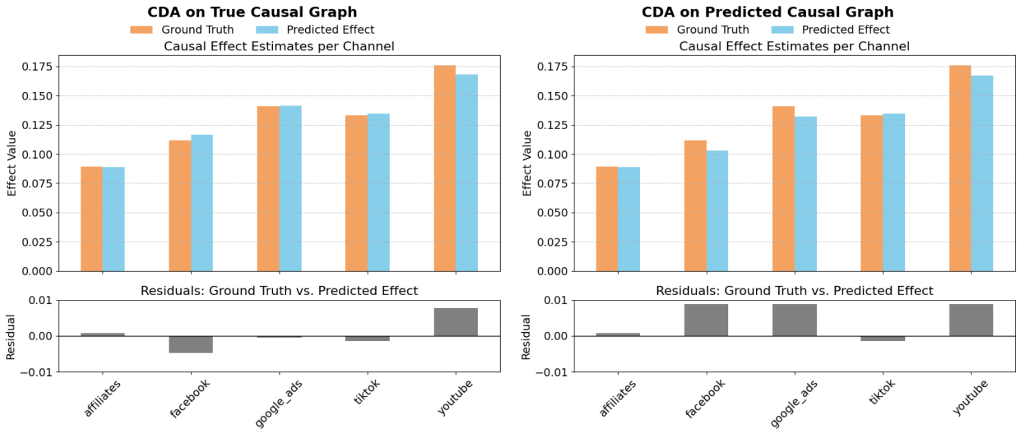
Figure 4.7: Comparison between the ground-truth DAG (a) and the DAG predicted using PCMCI (b).

Causal Effect Estimation To assess the performance of our causal effect estimator, we first evaluate it under ideal conditions by providing the true causal DAG as input. Since the simulated data are generated from this ground-truth structure, the estimated causal effects should closely match the true effects. This step verifies that the estimator is correctly specified and capable of recovering the underlying causal relationships when the structure is known. We then apply the estimation using the DAG learned by our discovery approach. To obtain robust estimates of causal effects, we follow the procedure described in Section 4.3.2 and compute CATE using a Monte Carlo approach. The estimation process is repeated 5,000 times under independent stochastic draws, and the resulting estimates are averaged to reduce variance.

It is important to note that small deviations are expected when using the true graph. Each channel inherently contains noise, which causes estimation error to accumulate along causal pathways with multiple parent nodes. For example, in the graph shown in Fig. 4.7 (a), both *Facebook* and *YouTube* have several outgoing connections, which increases noise propagation when the estimator calibrates their effects. As a result, channels with direct causal links such as *Affiliates*, *Google Ads*, and *TikTok* show nearly perfect recovery, whereas *Facebook* and *YouTube* display slightly higher residual error. This expectation is indeed reflected in the results shown in Fig. 4.8 (a).

When the estimator is applied to the predicted DAG (as depicted in Fig. 4.8 (b), the overall accuracy maintains a high level of accuracy, with effect estimates that closely reflect the true underlying causal effects across channels. As expected, the residuals increase slightly relative to the true-graph case, particularly for channels whose incoming or outgoing edges were misidentified in the learned structure. Nevertheless, the deviations are small, implying that the predicted graph is sufficiently accurate to support reliable causal effect estimation.

These observations are quantitatively supported by the results presented in Table 4.2. Using the true graph yields very low estimation error, with



(a) CDA result on True Causal Graph (a) CDA result on Predicted Causal Graph

Figure 4.8: Comparison of causal effect estimates and residuals for (a) the predicted causal graph and (b) the simulated ground-truth causal graph, compared to the true causal effects. Each panel shows the estimated causal effects per marketing channel and the residual differences between estimated and true values.

RRMSE of 3.16% and MAPE of 2.17%, along with perfect rank preservation as indicated by a Spearman correlation of 1.00. When the predicted graph is used, error measures increase moderately to RRMSE of 5.27% and MAPE of 4.20%, while rank correlation remains high at 0.90. This indicates that although structural imperfections in the learned graph introduce some additional estimation error, the overall magnitude of causal effects and their relative ordering across channels are still captured with high fidelity.

Table 4.2: Comparison of causal effect estimation accuracy using the true causal graph versus the predicted graph. Lower RRMSE and MAPE indicate better estimation accuracy, while higher Spearman correlation reflects better rank preservation.

DAG Input	RRMSE (%)	MAPE (%)	Spearman Corr
True Graph	3.16	2.17	1.00
Predicted Graph	5.27	4.20	0.90

In summary, the results indicate that the estimator performs effectively whether it receives the true DAG or the predicted DAG as input, confirming that the workflow is both reliable and robust. However, this evaluation is based on only a single synthetic instance. To obtain a more reliable and statistically grounded assessment, the next section presents a broader simulation study using the same evaluation metrics.

Aggregate Simulation Results

Table 4.3 summarizes the estimation performance across 1,000 simulations, generated by sampling 200 random DAGs for each depth level from 2 to 6 layers (corresponding to our five input channels and one target). This balanced design ensures that each layer of structural complexity contributes equally to the overall results. Several insights emerge regarding the behaviour and robustness of the causal effect estimator as DAG depth increases and the underlying causal architecture becomes more complex.

Table 4.3: Estimation performance across DAG depths. Values report mean $\pm$ standard deviation over 200 simulations per depth.

Input DAG	Metric	2 Layers	3 Layers	4 Layers	5 Layers	6 Layers	All
True	RRMSE (%)	0.79 ± 0.44	7.71 ± 7.36	10.48 ± 9.38	13.56 ± 10.81	14.96 ± 12.67	9.50 ± 10.44
	MAPE (%)	0.69 ± 0.38	4.43 ± 4.08	6.32 ± 5.80	9.00 ± 7.77	10.18 ± 9.00	6.12 ± 7.05
	Spearman ρ	0.98 ± 0.05	0.92 ± 0.16	0.87 ± 0.24	0.84 ± 0.22	0.84 ± 0.23	0.89 ± 0.20
Predicted	RRMSE (%)	12.92 ± 10.20	20.65 ± 12.97	25.50 ± 13.72	29.34 ± 15.37	32.76 ± 15.43	24.23 ± 15.31
	MAPE (%)	9.31 ± 7.38	15.97 ± 10.72	22.34 ± 26.91	24.52 ± 14.13	26.85 ± 14.07	19.80 ± 17.26
	Spearman ρ	0.78 ± 0.29	0.63 ± 0.38	0.50 ± 0.44	0.40 ± 0.47	0.37 ± 0.49	0.53 ± 0.45

Performance when using the true DAG Across different DAG depths, the estimator shows strong accuracy when provided with the true causal DAG. Both RRMSE and MAPE remain low for shallower graphs with two to three layers and increase gradually as depth grows, reflecting the inherent difficulty of recovering causal effects in structures with more parent–child dependencies and greater noise accumulation. Even in the deepest configurations with five to six layers, the estimation error remains moderate relative to the scale of the simulated effects. The Spearman rank correlation also remains consistently high, ranging from 0.84 to 0.98, indicating that the estimator reliably preserves the relative ordering of channel importance even as absolute estimation error increases with graph complexity.

Aggregated results across depths further confirm the estimator's robustness. With an overall RRMSE of 9.50%, MAPE of 6.12%, and Spearman correlation of 0.89, the estimator recovers true causal effects with high fidelity on average across a wide range of DAG structures. These metrics demonstrate strong performance under the ideal condition of perfect structural knowledge, establishing a benchmark against which estimates using the predicted graph can be evaluated. The consistent preservation of rank ordering is particularly relevant in practice, as marketing decision-makers often prioritize relative channel importance over precise point estimates of individual effects.

Performance when using the predicted DAG When the estimator relies on the predicted graph rather than the true DAG, estimation error increases, as expected. This decline in performance is statistically significant for deeper DAGs (five to six layers), where structural inaccuracies intensify the inherent complexity of the task. For instance, RRMSE grows sharply from 12.92% at depth 2 to 32.76% at depth 6, with MAPE exhibiting a similar increase from 9.31% to 26.85%. The Spearman correlation also drops steadily as the number of layers increases, falling from 0.78 at two layers to 0.37 at six layers, indicating reduced ability to correctly rank channel importance under structural misspecification.

Overall performance across simulations indicates that the estimator remains informative. With an RRMSE of 24.23% and MAPE of 19.80%, the estimated causal effects differ from the true magnitudes by roughly 20 to 25 percent on average. While this represents a non-trivial loss of numerical precision compared to using the true graph, it remains practically acceptable range for most real-world attribution scenarios, where measurement noise and structural uncertainty are unavoidable. The moderate Spearman ρ of 0.53 further indicates that the estimator retains meaningful discriminatory power, preserving the ability to distinguish high-impact channels from lower-impact ones, even if rankings are no longer perfectly reliable.

Taken together, these results show that the proposed causal-driven attribution approach remains useful under realistic conditions, providing meaningful insight into relative channel contributions despite substantial structural noise in the learned DAGs.

Effect of DAG depth Both the true-DAG and predicted-DAG results reveal a consistent pattern: estimation accuracy systematically declines as graph depth increases. This decline is intuitive, as deeper graphs introduce more intermediate nodes and pathways through which noise can propagate, amplifying variance in the effects. Additionally, deeper structures create more opportunities for small errors in estimated relationships to accumulate and magnify as they propagate downstream along the causal chain. Whether arising from incorrect edge magnitudes or missing or spurious connections in the learned graph, their cumulative impact increases substantially with each additional layer.

The impact of depth becomes particularly pronounced when the estimator relies on the predicted DAG rather than the true structure. In this scenario, inaccuracies in the learned graph interact with the inherent complexity of deeper causal architectures, resulting in compounding estimation errors that degrade both numerical precision and rank-ordering performance. This heightened deterioration underscores the sensitivity of causal effect estimation to structural misspecification, especially in settings where causal pathways span multiple layers or involve densely connected nodes.

Overall, the results show that while the estimator performs robustly for shallow graphs, deeper causal architectures pose substantially greater challenges. The degradation manifests both in terms of absolute numerical accuracy and in the model's ability to preserve the correct relative ordering of channel effects. In contexts of multi-layered indirect channel interactions, practitioners should expect greater uncertainty and place more emphasis on directional insights and relative rankings rather than precise point estimates. The results also highlight the value of pursuing accurate structural discovery, as even modest improvements in the learned graph can yield meaningful gains in downstream estimation quality when dealing with complex causal systems.

4.5 Discussion and Conclusion

More than ever, businesses rely on data-driven decision-making to allocate marketing resources and maximize return on investment. However, the growing integration of AI into data driven technologies in digital advertising, such as recommendation systems (Jabbar et al., 2020) and digital platforms (Purmonen et al., 2023), has made consumer journeys increasingly complex, creating multi touchpoint pathways that span numerous channels (Micheaux and Bosio, 2019; Hardcastle et al., 2025). Simultaneously, the rise of big data has introduced challenges related to missing and uncertain information, complicating the interpretation of marketing signals. Furthermore, stringent privacy regulations (as discussed in Section 4.1), together with platform-level restrictions such as the deprecation of third party cookies, have sharply reduced access to the user-level data required by traditional attribution models. Consequently, path-based attribution methods have become increasingly impractical, creating an urgent need for new approaches that operate in privacy-constrained environments while capturing the complex interdependencies of modern marketing ecosystems.

To the best of our knowledge this is the first causal-driven attribution utilising impression data without path information. When introducing a new attribution model, it is essential to consider the standards and challenges it must address, foremost among them the type of data required. As outlined in Section 4.2, many attribution methodologies provide reliable insights but depend critically on user-level path data which detailed sequences of individual user interactions across touchpoints are available. In the absence of such data, these methods cannot be applied, regardless of a firm's analytical capabilities. Our CDA framework addresses this limitation by operating on aggregated data that does not require tracking individual user journeys, while still capturing the relationships between marketing channels and conversions.

Second, our approach uses impressions as independent variables, representing a necessary evolution beyond path-based attribution models that rely primarily on clicks or visits. As privacy regulations increasingly restrict user-level tracking, organisations lose the ability to validate platform-reported click data, making the collection of accurate and reliable click records progressively more difficult. This shift aligns with growing industry recognition that impressions capture valuable brand exposure that may not trigger immediate interactions but can meaningfully influence downstream consumer behaviour over time. As discussed in Section 4.1, exposure effects often precede and facilitate subsequent conversions, even in the absence of direct click engagement. Accordingly, impression-based models not only address the practical limitations of diminished click tracking but also offer a more comprehensive view of marketing influence by capturing both direct response effects and longer-term awareness-building mechanisms that click-centric approaches overlook.

Third, the trustworthiness of attribution methods depends on their ability to uncover meaningful causal relationships among marketing channels rather than mere correlations (Hair Jr. and Sarstedt, 2021). Practitioners need to understand which channels truly drive outcomes, not simply which are statistically associated with conversions. Our CDA model addresses this need by estimating channel effects within a structured causal framework. Although many existing approaches provide reliable attribution, most rely on user-level path data (Section 4.2.3). While aggregated data preclude causal inference at the individual journey level, this limitation is not problematic in practice. Businesses make strategic budget decisions at the population level, where the primary concern is the overall influence of channels rather than the behaviour of specific individuals. Without a causal understanding of these effects, firms risk allocating resources to channels that appear effective but do not actually contribute to performance (Robertshaw, 2017). For this reason, our approach is designed to deliver credible causal inference from aggregated data, enabling actionable marketing decisions.

Finally, because our approach does not require personally identifiable information, it can be applied in environments where individual-level tracking is restricted by privacy or security regulations. This makes it particularly valuable for highly regulated industries such as pharmaceuticals, telecommunications, healthcare, and education, where data protection laws often render traditional tracking methods impractical or illegal (Seth and Ramakrishnan, 2025; Schultz and Brüggemann, 2025). Beyond these sectors, the framework provides a practical solution for organizations seeking effective attribution without incurring the substantial infrastructure and compliance costs associated with collecting, storing, and securing user-level data. By operating on aggregated data, our approach lowers both financial and regulatory barriers, making advanced attribution analysis accessible to organisations of varying sizes and capabilities.

4.5.1 Theoretical Implications

This paper contributes to the marketing and causal inference literature by proposing a novel causal-driven attribution framework designed to operate exclusively on aggregated marketing data, specifically impression-level time series, without relying on path-level or user-level tracking. This addresses a critical and growing need for privacy-preserving, scalable, and interpretable attribution models in marketing analytics, particularly in regulatory environments where access to granular, identifiable data is increasingly restricted or prohibited.

First, although SEM is traditionally used for path analysis and hypothesis testing (Stein et al., 2011), its generative nature makes it well suited for simulating synthetic data with controlled causal structure. In this study, we leverage SEM to generate datasets that follow predefined relationships among marketing channels and conversion outcomes. By explicitly encoding causal pathways, functional forms, and noise processes, SEM allows us to simulate realistic multivariate dependencies observed in marketing environments. While no simulation can fully replicate real behavioural data, SEM provides a coherent and interpretable approximation of the underlying data-generating process through its regression-based formulation (Iacobucci, 2009). These modelling assumptions offer a transparent, theoretically grounded foundation for controlled experimentation while preserving key features of causal structure, noise propagation, and channel interactions.

Second, our CDA framework reconceptualizes attribution as a causal inference problem rather than a correlational one, aligning with prior work (Smith and Pell, 2003; Hair Jr. and Sarstedt, 2021). It advances attribution theory by integrating causal discovery and causal estimation within a unified analytical pipeline. In the discovery stage, the framework employs the PCMCI algorithm to identify temporal conditional independencies and recover the directional structure among marketing channels using only impression-level time series, establishing a causal graph without requiring user-level tracking data. In the estimation stage, this structure is used to compute CATEs that capture both direct and indirect channel contributions to conversions, with a Monte Carlo procedure providing stable estimates under uncertainty. Together, these stages link structural discovery with causal effect quantification. Across multiple simulations, the framework recovers true causal effects with high fidelity (RRMSE 9.50%, Spearman correlation 0.89) when the true graph is known and maintains reasonable accuracy under the predicted graph (RRMSE 24.23%, Spearman correlation 0.53). Overall, this integration demonstrates that causal attribution can be achieved from aggregated data while preserving interpretability and robustness.

A third theoretical contribution of this study is the development of a comprehensive benchmarking framework for evaluating attribution models under controlled causal conditions. Since real-world marketing data rarely reveal the true causal relationships among channels, empirical validation of attribution models is typically constrained or indirect. By constructing a simulation environment grounded in SEM, predefined causal graphs, and systematically varied graph layers, our approach generates datasets where the true channel effects are known by design. This enables robust, repeatable comparisons of attribution performance across a wide range of structural complexities and data generating processes. Beyond evaluating our own framework, this benchmark provides a foundation for future research to assess alternative attribution methods under transparent and reproducible conditions. Thus, it contributes to the field's methodological foundations by establishing a standard approach for testing causal attribution models when true causal structure cannot be directly observed.

4.5.2 Managerial Implications

Our CDA model carries several practical implications for marketing managers, analysts, and decision-makers who operate in increasingly complex and privacy-constrained environments. First, managers no longer need to rely on user-level data or cookies to make data-driven decisions. Our approach applies only on aggregated impression-level data, making it highly applicable in industries where access to granular information is restricted by privacy (as detailed in Section 4.1). This opens the door for more organizations that can adopt evidence-based attribution strategies, including those that lack advanced data infrastructure.

A further managerial insight of this study concerns the ability of the causal-driven attribution framework to capture the full marketing funnel. Many traditional attribution models, including machine learning and deep learning approaches, emphasize immediate conversion signals by predicting whether specific sequences of user interactions lead to conversion and then inferring attribution from model weights or feature contributions (Filippou, Georgiadis and Jha, 2025). Such approaches often overlook the broader dynamics through which channels interact and influence one another over the customer journey. In contrast, our framework provides visibility into how digital channels jointly operate across the funnel, capturing both direct and indirect effects. This holistic perspective enables decision-makers to understand the collaborative role of channels, supporting more informed strategic planning and long-term business growth (Liu et al., 2022).

An important managerial contribution is related to marketing budget allocation. With global advertising spending projected to exceed $1.17 trillion in 2025 (Statista, 2025*b*), effective budget allocation has become a critical managerial priority. Existing literature presents several approaches to budget allocation (Zhao, Hua, Yan, Zhang, Xu and Yang, 2019; Fischer et al., 2011); however, most offer only partial visibility into the unique contributions of each channel within the overall marketing system. By estimating both the direct and indirect influence of each channel, our framework offers managers clearer and more actionable guidance for budget allocation. In addition, the ability to identify channels that exert negative or counterproductive influence enables firms to avoid wasteful spending and reallocate resources toward more effective activities.

Based on these insights, managers should reconsider how performance metrics and KPIs are evaluated within the marketing mix. Our method offers a more comprehensive view of channel influence, enabling refinement of existing measurement practices. This broader perspective helps managers rebalance media portfolios by identifying undervalued yet effective touchpoints that traditional models may overlook. The framework's simplicity and scalability make it suitable for both early-stage firms with limited analytical capacity and large organizations managing complex multichannel campaigns. By reducing data requirements while increasing analytical richness, our approach helps democratize marketing measurement and supports more informed strategic decision-making across diverse organizational contexts.

4.5.3 Limitations and Directions for Future Research

This study has several limitations that point to promising directions for future research. We begin with limitations related to causal discovery. The model relies only on observational data and does not incorporate interventional evidence, which is known to improve causal identification (Silva, 2016). In addition, as noted by Runge et al. (2019), causal inference depends on a set of underlying assumptions. Our approach assumes causal sufficiency, implying that all relevant variables are observed and that no unmeasured factors jointly influence multiple channels. It also relies on the causal Markov condition, whereby each variable is independent of its non-effects given its direct causes, and on faithfulness, which requires that observed conditional independencies reflect true structural independence rather than coincidental cancellations. Future work should explore integrating interventional data to relax these assumptions and improve robustness. Researchers should also enhance the simulation framework by incorporating more realistic temporal patterns and behavioural dynamics that better represent actual marketing environments.

Second, a wide range of causal discovery methodologies has been developed, as documented in the comprehensive review by Niu et al. (2024). Determining a suitable algorithm requires alignment between the analytical goals and the assumptions that can plausibly be applied to the data generating process. In this study, PCMCI was chosen because it is well suited to impression level time series and can accommodate temporal dependence structures that are common in marketing contexts. Nevertheless, other discovery techniques may achieve stronger performance under different modelling assumptions or where the data generating process follows structural patterns not well captured by PCMCI. The present work therefore provides a benchmark against which future research may evaluate and refine other causal discovery approaches, with the potential to enhance the accuracy and robustness of causal driven attribution models.

A further set of limitations arises from the causal estimation stage, which relies on observational data and thus cannot fully resolve the challenges of non-experimental inference. As Li and Kannan (2014) note, such models are inherently vulnerable to endogeneity and omitted variable bias. Although our framework accounts for temporal ordering and conditional relationships, unobserved heterogeneity or strategic behaviour, such as selective targeting or endogenous channel activation, may still distort the estimated causal effects. These challenges are amplified in marketing settings, where channels often exhibit interdependent spillover and carryover dynamics that are difficult to separate without controlled interventions. Future research could extend the present model by explicitly incorporating spillover and carryover mechanisms to better capture the interconnected nature of marketing activities.

Moreover, as Ferketich et al. (1984) discuss, residual analysis in causal models often reveals violations of key assumptions, including misspecified functional forms, nonlinearity, and correlated errors across equations. These violations can undermine causal validity, particularly when feedback mechanisms or dynamic dependencies exist but are not explicitly modelled. In our framework, residual dependence or patterns may signal unmodelled confounding or non-stationarity, potentially leading to biased causal effect estimates. Future work should thus explore robustness checks, alternative identification strategies, and experimental or quasi-experimental designs to validate the causal pathways identified through the proposed discovery and estimation stages.

5 Conclusion

5.1 Overview

This thesis is, at its core, a journey. It began not with a grand theoretical framework or a pre-determined methodological roadmap, but with a question rooted in genuine puzzlement. Prior to my research journey I was a marketing practitioner. While I was working in the travel industry, I got exposed to a particular consumer journey. Regardless of the marketign activity, users would visit the website or app to complete their booking via direct traffic source. In fact, a significant proportion of global online purchases are attributed to direct traffic source. A source that, by definition, can not be recorded as a marketing channel. As a result, similar problems faced by many others. This observation was difficult to accept. If direct traffic is where the conversion journey ends, it cannot plausibly be where it begins. Something must have driven those users to type a URL, click a bookmark, or return to a site they had visited before. The attribution systems in widespread use simply had no way of recording it.

That question, uncomfortable in its simplicity, set this research programme in motion. What follows is a summary of how each step emerged from the evidence and limitations of the one before it, and where the journey leads next.

5.2 The Journey: From Pattern Discovery to Causal Attribution

5.2.1 Step 1: Finding the Hidden Influence of Marketing Channels on Direct Traffic

Despite the volume of attribution research, certain segments of the consumer journey had gone largely unexamined. In particular, the relationship between observable channel-level traffic and the purchases attributed to direct traffic sources had received little systematic attention. Standard analytics platforms are designed to measure the isolated performance of each channel; they are not built to surface the interdependencies among them.

The first study addressed this gap directly. Using 89,394 purchase-level observations from a European e-commerce retailer, it applied linear regression with multicollinearity and autocorrelation diagnostics to test whether sessions originating from specific digital marketing channels were associated with conversion activity recorded under the direct traffic source. The results were unambiguous. Three channels: Google paid search, price comparison engines, and email marketing collectively explained 61% of the variance in direct traffic purchases. Channels that are evaluated solely on their own last-click contribution are, in reality, also driving a substantial share of the conversions that standard systems assign to a source with no visible marketing antecedent.

This finding established the empirical foundation of the thesis. It confirmed that cross-channel influence is not merely theoretically plausible but measurable on real data, and that the conventional attribution paradigm systematically undervalues the downstream contribution of performance-oriented channels.

5.2.2 Step 2: Uncovering the Brand Effects of Performance Campaigns

Having established that marketing channels influence outcomes beyond their own attributed conversions, the natural next question was whether a similar dynamic operated at the level of brand building. The prevailing view in digital media planning treats brand investment and performance investment as distinct activities, evaluated on separate metrics and managed with separate budgets. This separation is embedded in how agencies are structured, how campaigns are planned, and how success is reported.

The second study set out to test whether this separation holds empirically. Using weekly marketing investment and branded search data from a European e-commerce business operating across five countries, with an annual media budget exceeding 3 million, it applied ridge regression with adstock transformation to measure whether lower-funnel performance channels generate measurable effects on organic branded keyword searches. A recognised proxy for brand awareness and brand equity.

Every major performance channel was found to drive branded searches at a statistically significant level. Social media investment yielded the largest effect. The combined model raised explained variance from 65.8% to 76.8% once advertising carryover dynamics were properly accounted for.

These findings directly challenged the brand/performance dichotomy. They also began to reveal the limits of studying channels in pairs or in isolation: what the data showed, repeatedly, was a system of interdependencies. Understanding those interdependencies required a different kind of framework altogether.

5.2.3 Step 3: From Correlation to Causal Attribution

While the second study was in development, the regulatory environment for digital advertising was changing in ways that would make its core data requirements increasingly difficult to satisfy. The General Data Protection Regulation (GDPR) and the California Consumer Privacy Act (CCPA) had already materially restricted the ability of advertisers to collect and process individual-level behavioural data. Platform-level changes including Apple's App Tracking Transparency framework and the ongoing deprecation of third-party cookies, were further eroding the tracking infrastructure on which path-based attribution depends.

The implication was clear: models built around user-level journey data were becoming structurally impractical for a growing share of organisations. There was an urgent need for attribution approaches that could operate on the kind of aggregated, impression-level data that remained accessible under privacy-constrained conditions.

At the same time, the first two studies had established something important: channel interdependencies are real, they are directional, and they matter for measurement. Attribution models that ignore these interdependencies will systematically misrepresent channel value. What was needed was not just a privacy-compatible model, but a holistic attribution framework. One capable of discovering the structure of channel relationships without assuming it in advance, and then using that structure to produce causal effect estimates rather than correlational associations.

The third study developed exactly this. The Causal-Driven Attribution (CDA) framework is a two-stage pipeline: in the first stage, the PCMCI algorithm identifies temporal conditional independencies among marketing channel impression time series and recovers the directional causal structure among them, without requiring user-level data, click sequences, or path records. In the second stage, this discovered structure is used within a Structural Causal Model to estimate each channel's incremental contribution to conversions via Monte Carlo simulation of Conditional Average Treatment Effects.

But developing the framework raised an immediate problem. In real-world marketing data, the true causal contribution of each channel is unobservable. There is no ground truth against which to validate attribution estimates. To resolve this, the study developed a simulation benchmarking environment using Structural Equation Models with predefined causal parameters, enabling the generation of synthetic datasets in which the true channel effects are known by design. Across 1,000 simulations spanning five levels of causal graph complexity, the framework recovered true causal effects with an RRMSE of 9.50% and a Spearman rank correlation of 0.89 under ideal structural knowledge, and an RRMSE of 24.23% with a Spearman correlation of 0.53 when the causal structure was inferred entirely from data.

This confirmed that causal attribution is achievable from aggregated data alone, and that it is robust enough to be practically informative even when the underlying causal structure is not perfectly recovered.

5.3 Summary of Methods and Data

Table 5.1 summarises the three publications, the data employed in each, and the principal methods applied.

5.4 Collective Contribution

Taken together, the three studies constitute a coherent and progressive research programme. The first demonstrates, on real purchase data, that indirect channel influence is both measurable and substantial. The second demonstrates, across five markets, that the division between brand and performance is empirically untenable. The third translates these insights into a causal attribution framework that operates under the data constraints now facing the majority of practitioners.

Each study inherits its primary research question from the limitation of the one before it. The result is not three independent papers but a single argument built across three empirical and methodological stages: that the consumer journey is causally interconnected, that standard measurement systems are blind to most of that interconnectedness, and that causal inference methods applied to aggregated data can recover a meaningful portion of what conventional models miss — without requiring the erosion of user privacy as a precondition.

The contribution extends in four directions. Empirically, the thesis establishes cross-channel effects on real data at a scale and specificity not previously reported in this literature. Methodologically, it introduces the combination of ridge regression with adstock transformation for branded search modelling, and the CDA pipeline as the first impression-only causal attribution framework. Theoretically, it reconceptualises marketing attribution as a causal inference problem and reframes the consumer journey as a causally interconnected system. Practically, it provides tools that do not require specialised data infrastructure, making rigorous attribution accessible to the organisations. Particularly, small and medium-sized enterprises, that they need it the most.

5.5 Future Research Directions

This thesis opens several directions for future research, spanning methodological refinement, empirical extension, and the identification of new patterns in consumer behaviour.

5.5.1 Improving Causal Discovery Accuracy

The CDA framework relies on the PCMCI algorithm for causal structure recovery, and the simulation results show that estimation quality declines as graph depth increases. A natural priority for future work is improving the accuracy of the discovery stage. This could involve evaluating alternative causal discovery algorithms. Including constraint-based, score-based, and hybrid approaches against the benchmarking environment developed in this thesis. Additionally, incorporating interventional data alongside observational time series, where feasible, would relax the causal sufficiency assumption and improve identifiability in settings where unmeasured confounders are plausible.

The simulation framework itself could be extended to incorporate more realistic temporal dynamics, including seasonality, non-stationarity, and heterogeneous carryover structures, better approximating the data-generating processes of actual marketing environments.

5.5.2 Identifying Other Common Patterns of Consumer Behaviour

The first two studies surfaced two specific, previously underexamined patterns in the consumer journey: the downstream influence of performance channels on direct traffic conversions, and the brand awareness spillover generated by lower-funnel media investment. There is reason to believe that additional patterns of this type exist. Consumer journeys involve many more touchpoints, platforms, and behavioural transitions than the specific paths studied here.

Future research should systematically search for other recurring and undervalued patterns in how channels interact, how consumers transition between browsing and purchase states, and how impression exposure influences subsequent search and navigation behaviour. Cross-device and cross-platform dynamics, in particular, remain poorly mapped in the empirical attribution literature.

5.5.3 Making Attribution Decisions Without User-Level Data

Privacy regulation is not a transient constraint. The structural erosion of individual-level tracking data represents a permanent shift in the operating environment for marketing measurement. Future research should continue to develop the methodological infrastructure for effective attribution under these constraints.

Promising directions include the extension of the CDA framework to incorporate richer aggregated signals such as impression frequency distributions, reach curves, and media mix variation without reintroducing user-level identifiers. Bayesian approaches to causal discovery and estimation may offer advantages in data-sparse settings, providing principled uncertainty quantification alongside point estimates. Methods for combining aggregated observational data with the outputs of small-scale or historical experiments could further improve the reliability of causal effect estimates without requiring large-scale randomised trials.

More broadly, the field would benefit from shared benchmarking environments — along the lines of the synthetic data framework developed in Study 3 — that allow competing attribution methods to be evaluated under comparable and transparent conditions. Such infrastructure does not currently exist in a standardised form, and its absence is a material obstacle to cumulative methodological progress.

5.5.4 Scaling to Real-World Validation

The CDA framework has been validated on synthetic data. The next critical step is validation against empirical ground truth from controlled experiments geo-based holdout tests, incrementality studies, or matched-market designs. Such experiments are costly and operationally disruptive, but even a small number of real-world comparisons would substantially strengthen the evidence base for impression-level causal attribution and help calibrate the conditions under which the framework's accuracy is most and least reliable.

5.6 Closing Remarks

Understanding the consumer journey is one of the most consequential and most methodologically challenging problems in marketing science. A substantial body of research has advanced our knowledge of how consumers move across channels and touchpoints before converting. Yet the field continues to face a persistent tension: the most rigorous methods tend to be the least accessible, while the most widely adopted approaches often rest on assumptions that the empirical literature has repeatedly called into question.

This thesis engaged with that tension from both directions. Its methodological contribution lies in demonstrating that the analytical tools available to the field can be extended in ways that are both theoretically grounded and practically realisable. The combination of ridge regression with adstock transformation, the application of temporal causal discovery to aggregated marketing data, and the development of a simulation-based benchmarking environment each represent methodological steps that advance what is measurable, not merely what is measurable in principle. These contributions are intended to be of value to researchers seeking to study channel interdependencies, causal attribution, and privacy-constrained measurement under real-world conditions.

At the same time, the research was designed with an awareness that methodological rigour is necessary but not sufficient. The approach taken here was therefore deliberate in its structure: rather than beginning with a comprehensive framework, the thesis first isolated specific, underexamined patterns in the consumer journey and addressed them with methods transparent enough to be interpreted and replicated. Only once those empirical foundations were established did the research build toward a holistic causal attribution framework. The progression from targeted to holistic, from correlational to causal, was not incidental to the thesis it was the thesis.

The result is a body of work that aims to contribute to both the academic literature and to the broader project of making marketing measurement more rigorous, more honest about what conventional attribution systems miss, and more accessible to the organisations that need it most.

Table 5.1: Summary of publications, data, and methods.

Publication	**Data**	**Methods**	**Outcome**
Study 1 *Establishing the link: Does web traffic from various marketing channels influence direct traffic source purchases?* *Marketing Letters*, 2024	89,394 purchase-level observations from a European e-commerce retailer (EU, single market). Daily sessions by channel and direct traffic purchases.	Ordinary Least Squares (OLS) linear regression. Multicollinearity diagnostics (VIF). Autocorrelation diagnostics (Durbin–Watson). Outlier analysis.	3 channels (Google paid search, price comparison, email) explain 61% of variance in direct traffic purchases.
Study 2 *Unveiling digital dynamics: Do digital media investments impact organic branded searches?* *Journal of Marketing Analytics*, 2025	Weekly marketing investment and organic branded keyword search data from a European e-commerce business, 5 countries, annual media budget >3M.	Ridge regression (L2 regularisation). Adstock transformation (advertising carryover). Cross-validated regularisation parameter selection ($\lambda = 191.048$).	All major performance channels drive branded searches at a significant level. Social media yields the largest effect. R^2 improves from 0.658 to 0.768.
Study 3 *Causal-Driven Attribution (CDA): Estimating channel influence without user-level data* *arXiv:2512.21211*, 2025	Simulated aggregated impression-level time series (5 channels: Google Ads, Facebook, TikTok, YouTube, Affiliates). 1,000 synthetic datasets across DAG depths 2–6, generated via Structural Equation Models with predefined causal parameters.	PCMCI algorithm (temporal causal discovery). Structural Causal Model (DoWhy). Conditional Average Treatment Effects (CATE) via Monte Carlo simulation (5,000 draws). Benchmarking: RRMSE, MAPE, Spearman ρ, AUC, SHD, SID.	First impression-only causal attribution framework. RRMSE 9.50%, Spearman $\rho = 0.89$ (true graph). RRMSE 24.23%, Spearman $\rho = 0.53$ (predicted graph).

RRMSE = Relative Root Mean Squared Error; MAPE = Mean Absolute Percentage Error; SHD = Structural Hamming Distance; SID = Structural Intervention Distance; DAG = Directed Acyclic Graph; PCMCI = Peter and Clark Momentary Conditional Independence.

Bibliography

Akinwande, M., Dikko, H. and Samson, A. (2015), 'Variance inflation factor: as a condition for the inclusion of suppressor variable(s) in regression analysis', *Open Journal of Statistics* **5**, 754.

Alamsyah, A., Putri, F. and Sharif, O. (2014), Social network modeling approach for brand awareness, *in* '2nd International Conference on Information and Communication Technology (ICoICT)', IEEE, Bandung, Indonesia, pp. 448–453.

Anderl, E., Becker, I., von Wangenheim, F. and Schumann, J. (2016), 'Mapping the customer journey: Lessons learned from graph-based online attribution modeling', *International Journal of Research in Marketing* **33**, 457–474.

Anderl, E., Becker, I., Wangenheim, F. V. and Schumann, J. H. (2014), 'Mapping the customer journey: A graph-based framework for online attribution modeling', *Available at SSRN 2343077* .

Arava, S. K., Dong, C., Yan, Z., Pani, A. et al. (2018), 'Deep neural net with attention for multi-channel multi-touch attribution', *arXiv preprint arXiv:1809.02230* .

Archak, N., Mirrokni, V. S. and Muthukrishnan, S. (2010), Mining advertiser-specific user behavior using adfactors, *in* 'Proceedings of the 19th international conference on World wide web', pp. 31–40.

Asamoah, M. (2014), 'Re-examination of the limitations associated with correlational research', *Journal of Educational Research and Reviews* **2**, 45–52.

Awichanirost, J. and Phumchusri, N. (2020), Analyzing the effects of sessions on unique visitors and unique page views with Google Analytics: A case study of a tourism website in Thailand, *in* '2020 IEEE 7th International Conference on Industrial Engineering and Applications (ICIEA)', IEEE, Bangkok, Thailand, pp. 1014–1018.

Baltes, L. (2017), 'Marketing technology (Martech) – the most important dimension of online marketing', *Bulletin of the Transilvania University of Brasov, Series V: Economic Sciences* **10**, 43–48.

Ben Mrad, A. and Hnich, B. (2024), 'Intelligent attribution modeling for enhanced digital marketing performance', *Intelligent Systems with Applications* **21**, 200337.

Berman, R. (2018), 'Beyond the last touch: Attribution in online advertising', *Marketing Science* **37**(5), 771–792.

Bernarto, I., Berlianto, M., Meilani, Y., Masman, R. and Suryawan, I. (2020), 'The influence of brand awareness, brand image, and brand trust on brand loyalty', *Journal of Management* **24**, 412–426.

Bhargava, H., Rubel, O., Altman, E. et al. (2020), 'Platform data strategy', *Marketing Letters* **31**, 323–334.

Bianchi, T. (2022*a*), 'Global Website Traffic Distribution 2019, by Source'. [Accessed 26 May 2023].
URL: *https://www.statista.com/statistics/1110433/distribution-worldwide-website-traffic/*

Bianchi, T. (2022*b*), 'Share of Branded vs. Unbranded Search Traffic to Businesses 2018–2019'. [Accessed 19 August 2024].
URL: *https://www.statista.com/statistics/1143368/distribution-branded-search-volume-global-brands-smbs/*

Bilgic, E. and Duan, Y. (2019), E-commerce and business analytics: A literature review, *in* 'Digital Economy. Emerging Technologies and Business Innovation. Proceedings of the 4th International Conference, rt-PA 2019', Springer International Publishing, Beirut, Lebanon, pp. 173–182.

Bleier, A. and Eisenbeiss, M. (2015), 'Personalized online advertising effectiveness: The interplay of what, when, and where', *Marketing Science* **34**, 669–688.

Blozis, S. (2022), 'Bayesian two-part multilevel model for longitudinal media use data', *Journal of Marketing Analytics* **10**(4), 311–328.

Bonta, R. (2022*a*), 'California Consumer Privacy Act (CCPA)', *State of California Department of Justice* .

Bonta, R. (2022*b*), 'California consumer privacy act (ccpa)', *Retrieved from State of California Department of Justice: https://oag. ca. gov/privacy/ccpa* .

Breuer, R. and Brettel, M. (2012), 'Short-and long-term effects of online advertising: Differences between new and existing customers', *Journal of Interactive Marketing* **26**(3), 155–166.

Brinker, S. and Baldwin, J. (2020), *Martech 2030: Five Trends in Marketing Technology for the Decade of the Augmented Marketer*, Marketing Technology Media, LLC, Massachusetts, US.

Brookman, J., Rouge, P., Alva, A. and Yeung, C. (2017), 'Cross-device tracking: Measurement and disclosures', *Proceedings on Privacy Enhancing Technologies* **2017**, 133–148.

Bucklin, R. and Sismeiro, C. (2009), 'Click here for internet insight: Advances in clickstream data analysis in marketing', *Journal of Interactive Marketing* **23**(1), 35–48.

Buhalis, D. and Volchek, K. (2021), 'Bridging marketing theory and big data analytics: The taxonomy of marketing attribution', *International Journal of Information Management* **56**, 102253.

Cambra-Fierro, J., Polo-Redondo, Y. and Trifu, A. (2021), 'Short-term and long-term effects of touchpoints on customer perceptions', *Journal of Retailing and Consumer Services* **61**, 102520.

Chandrasekaran, D., Srinivasan, R. and Sihi, D. (2018), 'Effects of offline ad content on online brand search: Insights from Super Bowl advertising', *Journal of the Academy of Marketing Science* **46**, 403–430.

Chang, Y. and Thorson, E. (2004), 'Television and web advertising synergies', *Journal of Advertising* **33**, 75–84.

Chatfield, C. and Xing, H. (2019), *The Analysis of Time Series: An Introduction with R*, CRC Press, Boca Raton, London.

Chatterjee, R. S., Hameed, I. and Cham, T.-H. (2025), 'Cognitive and affective appraisal of online impulse buying: a multi-mediation approach', *Journal of Marketing Analytics* **13**, 1238–1257.

Chen, Y. (2016), 'Spatial autocorrelation approaches to testing residuals from least squares regression', *PLOS ONE* **11**, e0146865.

Cheng, L., Guo, R., Moraffah, R., Sheth, P., Candan, K. S. and Liu, H. (2022), 'Evaluation methods and measures for causal learning algorithms', *IEEE Transactions on Artificial Intelligence* **3**(6), 924–943.

Chuang, Y. and Thomson, K. (2017), 'Maneuvering multimarket competition: The effects of multimarket contact and strategic alliances on performance of single-market firms', *Strategic Organization* **15**(3), 390–409.

Cornell, C., Mitchell, L. and Roughan, M. (2024), 'Rank is all you need: development and analysis of robust causal networks', *Applied Network Science* **9**(1), 39.

Cramer-Flood, E. (2023), 'Worldwide Digital Ad Spending 2023'. [Accessed 02 December 2022].

URL: *https://www.insiderintelligence.com/content/worldwide-digital-ad-spending-2023*

Cui, T., Ghose, A., Halaburda, H., Iyengar, R., Pauwels, K., Sriram, S., Tucker, C. and Venkataraman, S. (2021), 'Informational challenges in omnichannel marketing: Remedies and future research', *Journal of Marketing* **85**, 103–120.

Cyert, R. M. and March, J. G. (2020), *A behavioral theory of the firm*, Barakaldo Books.

Dalessandro, B., Perlich, C., Stitelman, O. and Provost, F. (2012), Causally motivated attribution for online advertising, *in* 'Proceedings of the Sixth International Workshop on Data Mining for Online Advertising and Internet Economy', Association for Computing Machinery, Beijing, China.

Danaher, P. and van Heerde, H. (2018), 'Delusion in attribution: Caveats in using attribution for multimedia budget allocation', *Journal of Marketing Research* **55**, 667–685.

Dens, N. and Poels, K. (2023), 'The rise, growth, and future of branded content in the digital media landscape', *International Journal of Advertising* **42**(1), 141–150.

Despotovic, M., Nedic, V., Despotovic, D. and Cvetanovic, S. (2016), 'Evaluation of empirical models for predicting monthly mean horizontal diffuse solar radiation', *Renewable and Sustainable Energy Reviews* **56**, 246–260.

Devaux, R. and Bomsel, O. (2021), 'Cross-Media Externalities in Advertising Markets'. CERNA Working Paper.
URL: *https://www.cerna.minesparis.psl.eu/Donnees/data17/1796-* $Papier_R D_C ross - media.pdf$

Dhebar, A. (2013), 'Toward a compelling customer touchpoint architecture', *Business Horizons* **56**, 199–205.

Donndelinger, J. A. and Ferguson, S. M. (2020), 'Design for the marketing mix: The past, present, and future of market-driven engineering design', *Journal of Mechanical Design* **142**(6), 060801.

Drivas, I., Sakas, D. and Giannakopoulos, G. (2021), Display advertising and brand awareness in search engines: Predicting the engagement of branded search traffic visitors, *in* 'Business Intelligence and Modelling in Entrepreneurship: Unified Approach with Simulation and Strategic Modelling, 8th IC-BIM 2019', Springer International Publishing, Cham, pp. 3–15.

Dufour, J., Gaudry, M. and Liem, T. (1980), 'The Cochrane-Orcutt procedure numerical examples of multiple admissible minima', *Economics Letters* **6**, 43–48.

Edell, J. and Keller, K. (1989), 'The information processing of coordinated media campaigns', *Journal of Marketing Research* **26**, 149–163.

Edelman, D. (2010), 'Branding in the Digital Age: You're Spending Your Money in All the Wrong Places'. [Accessed 12 May 2023].
URL: *https://hbr.org/2010/12/branding-in-the-digital-age-youre-spending-your-money-in-all-the-wrong-places*

El Mekkaoui, Z. E. A. and Benyoussef, H. (2024), The evolution of multi-touch attribution: Integrating advanced technologies for superior marketing insights, *in* 'International Conference on Advanced Intelligent Systems for Sustainable Development', Springer, pp. 426–435.

eMarketer (2020), 'Global Digital Ad Spending 2019'. [Accessed 10 September 2024].
URL: *https://www.emarketer.com/content/global-digital-ad-spending-2019*

Ephron, E. and McDonald, C. (2002), 'Media scheduling and carry-over effects', *Journal of Advertising Research* **42**, 66–70.

European Commission (2023), 'Annual Report on European SMEs 2022/2023'.

Ferketich, S. L., Verran, J. A., Hefferin, E. A., McLaughlin, F. E., Smith, R. A., Hinshaw, A. S. and Young, K. J. (1984), 'Residual analysis for causal model assumptions', *Western Journal of Nursing Research* **6**(1), 41–76.

Filippou, G., Georgiadis, A. G. and Jha, A. K. (2024), 'Establishing the link: Does web traffic from various marketing channels influence direct traffic source purchases?', *Marketing Letters* **35**(1), 59–71.

Filippou, G., Georgiadis, A. G. and Jha, A. K. (2025), 'Unveiling digital dynamics: Do digital media investments impact organic branded searches?', *Journal of Marketing Analytics* .

Filippou, G., Quach, B. M., Lenghel, D., White, A. and Jha, A. K. (2025), 'Causal-driven attribution (CDA): Estimating channel influence without user-level data', *arXiv preprint arXiv:2512.21211* .

Fischer, M., Albers, S., Wagner, N. and Frie, M. (2011), 'Practice prize winner—dynamic marketing budget allocation across countries, products, and marketing activities', *Marketing Science* **30**(4), 568–585.

Furquim, T., Da Veiga, C., Veiga, C. and Silva, W. (2022), 'The different phases of the omnichannel consumer buying journey: A systematic literature review and future research directions', *Journal of Theoretical and Applied Electronic Commerce Research* **18**(1), 79–104.

Gaur, J. and Bharti, K. (2020), 'Attribution modelling in marketing: Literature review and research agenda', *Academy of Marketing Studies Journal* **24**(4), 1–21.

Gaur, J., Bharti, K. and Bajaj, R. (2024), 'Maximizing marketing impact: Heuristic vs ensemble models for attribution modeling', *Global Knowledge, Memory and Communication* .

Ghose, A. and Todri-Adamopoulos, V. (2016), 'Toward a digital attribution model', *MIS Quarterly* **40**(4), 889–910.

Gijsenberg, M., Van Heerde, H., Dekimpe, M. and Nijs, V. (2011), 'Understanding the Role of Adstock in Advertising Decisions'. SSRN Working

Paper.
URL: *http://doi.org/10.2139/ssrn.1905426*

Google Analytics Help Center (2023), 'About the Attribution Model Comparison Tool'. [Accessed 23 April 2023].
URL: *https://support.google.com/analytics/answer/1662518*

Gupta, A., Su, B. and Walter, Z. (2004), 'An empirical study of consumer switching from traditional to electronic channels: A purchase-decision process perspective', *International Journal of Electronic Communications* **8**, 131–161.

Ha, H. and Perks, H. (2005), 'Effects of consumer perceptions of brand experience on the web: Brand familiarity, satisfaction and brand trust', *Journal of Consumer Behavior* **4**, 438–452.

Hair Jr., J. F. and Sarstedt, M. (2021), 'Data, measurement, and causal inferences in machine learning: Opportunities and challenges for marketing', *Journal of Marketing Theory and Practice* **29**(1), 65–77.

Hamilton, R. and Price, L. (2019), 'Consumer journeys: Developing consumer-based strategy', *Journal of the Academy of Marketing Science* **47**, 187–191.

Hardcastle, K., Vorster, L. and Brown, D. M. (2025), 'Understanding customer responses to ai-driven personalized journeys: impacts on the customer experience', *Journal of Advertising* **54**(2), 176–195.

Havlena, W., Cardarelli, R. and De Montigny, M. (2007), 'Quantifying the isolated and synergistic effects of exposure frequency for TV, print, and internet advertising', *Journal of Advertising Research* **47**, 215–221.

Heider, F. (2013), *The psychology of interpersonal relations*, Psychology Press.

Herman, R. J., Wahl, J., Ninad, U. and Runge, J. (2025), Unitless unrestricted Markov-consistent SCM generation: Better benchmark datasets for causal discovery, *in* 'Causal Learning and Reasoning', PMLR, pp. 1506–1531.

Hodson, T. (2022), 'Root-mean-square error (RMSE) or mean absolute error (MAE): When to use them or not', *Geoscientific Model Development* **15**, 5481–5487.

Hoerl, A. and Kennard, R. (1970), 'Ridge regression: Biased estimation for nonorthogonal problems', *Technometrics* **12**, 55–67.

Holzwarth, M., Janiszewski, C. and Neumann, M. (2006), 'The influence of avatars on online consumer shopping behavior', *Journal of Marketing* **70**, 19–36.

Huang, S. and Lin, Y. (2022), 'Exploring consumer online purchase and search behavior: An FCB grid perspective', *Asia Pacific Management Review* **27**, 245–256.

Hui, S., Fader, P. and Bradlow, E. (2009), 'Path data in marketing: An integrative framework and prospectus for model building', *Marketing Science* **28**(2), 320–335.

Hyndman, R. and Athanasopoulos, G. (2018), *Forecasting: Principles and Practice*, OTexts, Australia.

Iacobucci, D. (2009), 'Everything you always wanted to know about SEM (structural equations modeling) but were afraid to ask', *Journal of Consumer Psychology* **19**(4), 673–680.

Imbens, G. W. and Rubin, D. B. (2015), *Causal Inference in Statistics, Social, and Biomedical Sciences*, Cambridge University Press.

Jabbar, A., Akhtar, P. and Dani, S. (2020), 'Real-time big data processing for instantaneous marketing decisions: A problematization approach', *Industrial Marketing Management* **90**, 558–569.

James, G., Witten, D., Hastie, T. and Tibshirani, R. (2013), *An Introduction to Statistical Learning*, Vol. 112, Springer.

Jamieson, P. D., Porter, J. R. and Wilson, D. (1991), 'A test of the computer simulation model arcwheat1 on wheat crops grown in new zealand', *Field crops research* **27**(4), 337–350.

Ji, W. and Wang, X. (2017), Additional multi-touch attribution for online advertising, *in* 'Proceedings of the aaai conference on artificial intelligence', Vol. 31.

Jin, Y., Wang, Y., Sun, Y., Chan, D. and Koehler, J. (2017), Bayesian Methods for Media Mix Modeling With Carryover and Shape Effects, Technical report, Google Inc., New York.

Job, S., Tao, X., Cai, T., Xie, H., Li, L., Li, Q. and Yong, J. (2025), 'Exploring causal learning through graph neural networks: An in-depth review', *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* **15**(2), e70024.

Johnson, G. (2023), 'Inferno: A guide to field experiments in online display advertising', *Journal of Economics and Management Strategy* **32**, 469–490.

Joo, M., Wilbur, K., Cowgill, B. and Zhu, Y. (2014), 'Television advertising and online search', *Management Science* **60**, 56–73.

Joo, M., Wilbur, K. and Zhu, Y. (2016), 'Effects of TV advertising on keyword search', *International Journal of Research in Marketing* **33**(3), 508–523.

Joseph, J. (2006), 'Understanding Advertising Adstock Transformations'. SSRN Working Paper.
URL: *http://doi.org/10.2139/ssrn.924128*

Kadyrov, T. and Ignatov, D. (2019), Attribution of customers' actions based on machine learning approach?, *in* 'CEUR Workshop Proceedings', pp. 77–88.

Kakalejčík, L., Bucko, J., Resende, P. and Ferencova, M. (2018), 'Multi-channel marketing attribution using markov chains', *Journal of Applied Management and Investments* **7**(1), 49–60.

Kakalejčík, L., Bucko, J. and Danko, J. (2020), 'Impact of direct traffic effect on online sales', *Journal of Research in Interactive Marketing* **14**, 17–32.

Kakalejčík, L., Bucko, J., Resende, P. and Ferencova, M. (2018), 'Multi-channel marketing attribution using Markov chains', *Journal of Applied Management and Investments* **7**(1), 49–60.

Kamena, R. (2021), 'Moving towards inferential attribution modelling in a world without third-party cookies', *Applied Marketing Analytics* **7**, 122–130.

Kannan, P., Reinartz, W. and Verhoef, P. (2016), 'The path to purchase and attribution modeling: Introduction to special section', *International Journal of Research in Marketing* **33**, 449–456.

Kelley, H. H. (1967), Attribution theory in social psychology., *in* 'Nebraska symposium on motivation', University of Nebraska Press.

Kelly, J., Vaver, J. and Koehler, J. (2018), A causal framework for digital attribution, Technical report, Technical report, Google LLC, 2018. URL https://ai. google/research/pubs . . . .

Kireyev, P., Pauwels, K. and Gupta, S. (2016), 'Do display ads influence search? Attribution and dynamics in online advertising', *International Journal of Research in Marketing* **33**, 475–490.

Kitts, B., Wei, L., Au, D., Powter, A. and Burdick, B. (2010), Attribution of conversion events to multi-channel media, *in* '2010 IEEE International Conference on Data Mining', IEEE, Sydney, NSW, Australia, pp. 881–886.

Kotler, P., Keller, K., Brady, M., Goodman, M. and Hansen, T. (2016), *Marketing Management*, 3rd european edn, Pearson.

Kufile, O. T., Akinrinoye, O. V., Onifade, A. Y., Ejike, O. G., Otokiti, B. O. and Umezurike, S. A. (2023), 'Developing conceptual attribution models for cross-platform marketing performance evaluation', *International Journal of Multidisciplinary Research and Growth Evaluation* **4**(2), 844–854.

Kumar, S., Gupta, G., Prasad, R., Chatterjee, A., Vig, L. and Shroff, G. (2020), Camta: Causal attention model for multi-touch attribution, *in* '2020 international conference on data mining workshops (icdmw)', IEEE, pp. 79–86.

Kvíčala, D., Králová, M. and Suchánek, P. (2024), 'The impact of online purchase behaviour on customer lifetime value', *Journal of Marketing Analytics* .

LaBrecque, A. C., Voorhees, C. M., Khodakarami, F. and Fombelle, P. W. (2024), 'Native advertising effectiveness: The role of congruence and consumer annoyance on clicks, bounces, and visits', *Journal of the Academy of Marketing Science* **52**(6), 1692–1712.

Lambrecht, A. and Tucker, C. (2013), 'When does retargeting work? Information specificity in online advertising', *Journal of Marketing Research* **50**, 561–576.

Leeflang, P., Verhoef, P., Dahlström, P. and Freundt, T. (2014), 'Challenges and solutions for marketing in a digital era', *European Management Journal* **32**(1), 1–12.

Leeflang, P. and Wittink, D. (2000), 'Building models for marketing decisions: Past, present and future', *International Journal of Research in Marketing* **17**, 105–126.

Lemon, K. and Verhoef, P. (2016), 'Understanding customer experience throughout the customer journey', *Journal of Marketing* **80**(6), 69–96.

Lewis, R., Zettelmeyer, F., Gordon, B. R., Garib, C., Hermle, J., Perry, M., Romero, H. and Schnaidt, G. (2025), Amazon ads multi-touch attribution, Technical report.

Li, H. and Kannan, P. (2014), 'Attributing conversions in a multichannel online marketing environment: An empirical model and a field experiment', *Journal of Marketing Research* **51**, 40–56.

Li, H., Yu, L. and He, W. (2019*a*), 'The impact of GDPR on global technology development'.

Li, H., Yu, L. and He, W. (2019*b*), 'The impact of gdpr on global technology development'.

Liu, W., Liang, Y., Tang, O. and Ma, X. (2022), 'Channel competition and collaboration in the presence of hybrid retailing', *Transportation Research Part E: Logistics and Transportation Review* **160**, 102658.

Lovett, M. and Staelin, R. (2016), 'The role of paid, earned, and owned media in building entertainment brands: Reminding, informing, and enhancing enjoyment', *Marketing Science* **35**, 142–157.

Ma, N., Zhao, S., Sun, Z., Wu, X. and Zhai, Y. (2019), 'An improved ridge regression algorithm and its application in predicting TV ratings', *Multimedia Tools and Applications* **78**, 525–536.

Malter, M., Holbrook, M., Kahn, B., Parker, J. and Lehmann, D. (2020), 'The past, present, and future of consumer research', *Marketing Letters* **31**, 137–149.

Manjunath, C., Padigar, M. and Pedada, K. (2024), 'The role of digital orientation and strategic emphasis in creating brand competitiveness', *Journal of Retailing and Consumer Services* **80**, 103906.

March, J. G. and Simon, H. A. (1993), *Organizations*, John wiley & sons.

McDonald, G. (2009), 'Ridge regression', *WIREs Computational Statistics* **1**, 93–100.

McDonald, G. (2010), 'Tracing ridge regression coefficients', *WIREs Computational Statistics* **2**, 695–703.

McGuirk, M. (2023), 'Performing web analytics with Google Analytics 4: A platform review', *Journal of Marketing Analytics* **11**(4), 854–868.

Micheaux, A. and Bosio, B. (2019), 'Customer journey mapping as a new way to teach data-driven marketing as a service', *Journal of Marketing Education* **41**(2), 127–140.

Milano, R., Baggio, R. and Piattelli, R. (2011), The effects of online social media on tourism websites, *in* R. Law, M. Fuchs and F. Ricci, eds, 'Information and Communication Technologies in Tourism 2011', Springer Vienna, Vienna, pp. 471–483.

Moe, W. and Fader, P. (2004), 'Dynamic conversion behavior at E-commerce sites', *Management Science* **50**, 326–335.

Morgan, N. A., Clark, B. H. and Gooner, R. (2002), 'Marketing productivity, marketing audits, and systems for marketing performance assessment: Integrating multiple perspectives', *Journal of Business Research* **55**(5), 363–375.

Naik, P. and Peters, K. (2009), 'A hierarchical marketing communications model of online and offline media synergies', *Journal of Interactive Marketing* **23**, 288–299.

Naik, P. and Raman, K. (2003), 'Understanding the impact of synergy in multimedia communications', *Journal of Marketing Research* **40**, 375–388.

Nam, K. and Cho, K. (2025), 'Navigating the digital purchase journey: the impact of device type and visit channels on consumer behavior', *Industrial Management & Data Systems* .

Nashed, S. B., Mahmud, S., Goldman, C. V. and Zilberstein, S. (2025), 'Causal explanations for sequential decision making', *Journal of Artificial Intelligence Research* **83**.

Neslin, S. and Shankar, V. (2009), 'Key issues in multichannel customer management: Current knowledge and future directions', *Journal of Interactive Marketing* **23**, 70–81.

Niu, W., Gao, Z., Song, L. and Li, L. (2024), 'Comprehensive review and empirical evaluation of causal discovery algorithms for numerical data', *arXiv preprint arXiv:2407.13054* .

OECD (2019), *Unpacking E-Commerce: Business Models, Trends and Policies*, OECD Publishing, Paris.

Oh, T., Keller, K., Neslin, S., Reibstein, D. and Lehmann, D. (2020), 'The past, present, and future of brand research', *Marketing Letters* **31**, 151–162.

Oksanen, T. (2022), 'Companies' maturity for the deprecation of third-party cookies'.

Omidvar, M., Mirabi, V. and Shokry, N. (2011), 'Analyzing the impact of visitors on page views with Google Analytics', *International Journal of Web & Semantic Technology* **2**(1).

Onifade, A. Y., Ogeawuchi, J. C., Abayomi, A., Agboola, O. and George, O. (2021), 'Advances in multi-channel attribution modeling for enhancing marketing ROI in emerging economies', *Iconic Research and Engineering Journals* **5**(6), 360–376.

Papapetrou, P., Gionis, A. and Mannila, H. (2011), A Shapley value approach for influence attribution, *in* 'Joint European Conference on Machine Learning and Knowledge Discovery in Databases', Springer, pp. 549–564.

Pauwels, K., Currim, I., Dekimpe, M., Hanssens, D., Mizik, N., Ghysels, E. and Naik, P. (2004), 'Modeling marketing dynamics by time series econometrics', *Marketing Letters* **15**, 167–183.

Pearl, J. (2009*a*), *Causality*, Cambridge University Press.

Pearl, J. (2009*b*), *Causality*, Cambridge university press.

Pfeifer, P. E. and Carraway, R. L. (2000), 'Modeling customer relationships as markov chains', *Journal of interactive marketing* **14**(2), 43–55.

Plaza, B. (2011), 'Google Analytics for measuring website performance', *Tourism Management* **32**, 477–481.

Polat, E. and Gunay, S. (2015), 'The comparison of partial least squares regression, principal component regression and ridge regression with multiple linear regression for predicting PM10 concentration level based on meteorological parameters', *Journal of Data Science* **13**, 663–692.

Poutanen, R. (2020), Analysis of Online Advertisement Performance Using Markov Chains, Master's thesis, University of Tampere, Tampere, Finland.

Purmonen, A., Jaakkola, E. and Terho, H. (2023), 'B2b customer journeys: Conceptualization and an integrative framework', *Industrial marketing management* **113**, 74–87.

Qin, V., Pauwels, K. and Zhou, B. (2024), 'Data-driven budget allocation of retail media by ad product, funnel metric, and brand size', *Journal of Marketing Analytics* pp. 1–15.

Raman, K., Mantrala, M. K., Sridhar, S. and Tang, Y. (2012), 'Optimal resource allocation with time-varying marketing effectiveness, margins and costs', *Journal of Interactive Marketing* **26**(1), 43–52.

Reiley, D., Lewis, R. A. and Schreiner, T. (2012), 'Ad attributes and attribution: Large-scale field experiments measure online customer acquisition', *Available at SSRN 2049457* .

Ren, K., Fang, Y., Zhang, W., Liu, S., Li, J., Zhang, Y., Yu, Y. and Wang, J. (2018), Learning multi-touch conversion attribution with dual-attention mechanisms for online advertising, *in* 'Proceedings of the 27th ACM International Conference on Information and Knowledge Management', pp. 1433–1442.

Ribeiro, L. and Maués, A. (2023), Traditional brands vs. born-digital brands: The case of the fashion industry, *in* ‘Advances in Digital Marketing & eCommerce Conference’, Springer Nature, Cham, pp. 298–304.

Robertshaw, G. S. (2017), ‘Measuring media channel performance: a proposed alternative to the ‘last-in wins’ methodology’, *Journal of Marketing Analytics* **5**, 121–130.

Romero Leguina, J., Cuevas Rumín, Á. and Cuevas Rumín, R. (2020), ‘Digital marketing attribution: Understanding the user path’, *Electronics* **9**, 1822.

Runge, J., Bathiany, S., Bollt, E., Camps-Valls, G., Coumou, D., Deyle, E., Glymour, C., Kretschmer, M., Mahecha, M. D., Muñoz-Marí, J. et al. (2019), ‘Inferring causation from time series in earth system sciences’, *Nature communications* **10**(1), 2553.

Rust, R. T., Ambler, T., Carpenter, G. S., Kumar, V. and Srivastava, R. K. (2004), ‘Measuring marketing productivity: Current knowledge and future directions’, *Journal of Marketing* **68**(4), 76–89.

Rutz, O., Trusov, M. and Bucklin, R. (2011), ‘Modeling indirect effects of paid search advertising: Which keywords lead to more future visits?’, *Marketing Science* **30**, 646–665.

Sandeep, N., Vasavi, S. and Shruthi, G. (2021*a*), ‘Channel attribution using Markov chain model’, *Turkish Online Journal of Qualitative Inquiry* **12**(7).

Sandeep, N., Vasavi, S. and Shruthi, G. (2021*b*), ‘Channel attribution using markov chain model.’, *Turkish Online Journal of Qualitative Inquiry* **12**(7).

Santos, S. and Gonçalves, H. M. (2024), ‘Consumer decision journey: Mapping with real-time longitudinal online and offline touchpoint data’, *European Management Journal* **42**(3), 397–413.

Santos, S. and Gonçalves, H. (2021), 'The consumer decision journey: A literature review of the foundational models and theories and a future perspective', *Technological Forecasting and Social Change* **173**, 121117.

Sarasvuo, S., Rindell, A. and Kovalchuk, M. (2022), 'Toward a conceptual understanding of co-creation in branding', *Journal of Business Research* **139**, 543–563.

Schultz, C. D. and Brüggemann, P. (2025), 'The role of privacy concerns and technology anxiety regarding digital voice assistants for brand-and price-conscious grocery shoppers', *Journal of Marketing Analytics* .

Schweidel, D. A., Bart, Y., Inman, J. J., Stephen, A. T., Libai, B., Andrews, M., Rosario, A. B., Chae, I., Chen, Z., Kupor, D. et al. (2022), 'How consumer digital signals are reshaping the customer journey', *Journal of the Academy of Marketing Science* **50**(6), 1257–1276.

Schweidel, D. A., Bart, Y., Inman, J. J., Stephen, A. T., Libai, B., Andrews, M. et al. (2022), 'How consumer digital signals are reshaping the customer journey', *Journal of the Academy of Marketing Science* **50**(6), 1257–1276.

Seth, S. and Ramakrishnan, S. (2025), Innovations in multi-touch attribution models: Beyond traditional marketing analytics in the pharmaceutical sector, *in* '2025 International Conference on Emerging Smart Computing and Informatics (ESCI)', IEEE, pp. 1–5.

Shankar, V. and Kushwaha, T. (2021), 'Omnichannel marketing: Are cross-channel effects symmetric?', *International Journal of Research in Marketing* **38**, 290–310.

Shao, X. and Li, L. (2011), Data-driven multi-touch attribution models, *in* 'Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining', Association for Computing Machinery, San Diego, California, USA, pp. 258–264.

Shapley, L. S. et al. (1953), *A value for n-person games*, Princeton University Press Princeton.

Shavitt, S. and Barnes, A. (2020), 'Culture and the consumer journey', *Journal of Retailing* **96**, 40–54.

Sheather, S. (2009), *A Modern Approach to Regression with R*, Springer Texts in Statistics, Springer, New York, NY.

Shojaie, A. and Fox, E. B. (2022), 'Granger causality: A review and recent advances', *Annual Review of Statistics and Its Application* **9**(1), 289–319.

Silva, R. (2016), 'Observational-interventional priors for dose-response learning', *Advances in Neural Information Processing Systems* **29**.

Simon, H. A. (1955), 'A behavioral model of rational choice', *The quarterly journal of economics* pp. 99–118.

Simon, J. P. (2021), 'Apis, the glue under the hood. looking for the "api economy"', *Digital Policy, Regulation and Governance* **23**(5), 489–508.

Simon, M. and Goes, J. (2013), Assumptions, Limitations, Delimitations, and Scope of the Study, PhD thesis, Dissertation Success, LLC, Seattle, WA.

Singh, C. P. and Yousuf, R. (2024), Enhancing marketing strategies through big data-driven customer journey mapping: An analysis using machine learning algorithms, *in* '2024 International Conference on Emerging Innovations and Advanced Computing (INNOCOMP)', IEEE, pp. 479–484.

Sirdeshmukh, D., Ahmad, N., Khan, M. and Ashill, N. (2018), 'Drivers of user loyalty intention and commitment to a search engine: An exploratory study', *Journal of Retailing and Consumer Services* **44**, 71–81.

Skow, J. (2023), 'What is Direct Traffic in Google Analytics? (2023 Beginner's Guide)'. [Accessed 27 May 2023].
URL: *https://www.monsterinsights.com/what-is-direct-traffic-in-google-analytics/*

Smith, G. C. and Pell, J. P. (2003), 'Parachute use to prevent death and major trauma related to gravitational challenge: systematic review of randomised controlled trials', *Bmj* **327**(7429), 1459–1461.

Sokol, D. D. and Zhu, F. (2021), 'Harming competition and consumers under the guise of protecting privacy: An analysis of Apple's iOS 14 policy updates', *Cornell Law Review Online* **107**, 94.

Sokolova, N. and Titova, O. (2019), Digital marketing as a type: Concept, tools and effects, *in* 'Proceedings of the 1st International Scientific Conference 'Modern Management Trends and the Digital Economy: From Regional Development to Global Economic Growth' (MTDE 2019)', Atlantis Press, pp. 508–512.

Solcansky, M. and Simberova, I. (2010), 'Measurement of marketing effectiveness', *Economics and management* **15**, 755–759.

Srinivasan, V., Park, C. and Chang, D. (2005), 'An approach to the measurement, analysis, and prediction of brand equity and its sources', *Management Science* **51**, 1433–1448.

Stankevich, A. (2017), 'Explaining the consumer decision-making process: Critical literature review', *Journal of International Business Research and Marketing* **2**(6).

Starov, O., Zhou, Y., Zhang, X., Miramirkhani, N. and Nikiforakis, N. (2018), Betrayed by your dashboard: Discovering malicious campaigns via web analytics, *in* 'Proceedings of the 2018 World Wide Web Conference', pp. 227–236.

Statista (2025*a*), 'Advertising Worldwide – Statistics & Forecasts'. [Accessed 28 October 2025].
URL: *https://www.statista.com/outlook/amo/advertising/worldwide*

Statista (2025*b*), 'Advertising worldwide – statistics & forecasts', https://www.statista.com/outlook/amo/advertising/worldwide. Accessed: 28 October 2025.

Stein, C. M., Morris, N. J. and Nock, N. L. (2011), Structural equation modeling, *in* 'Statistical human genetics: Methods and protocols', Springer, pp. 495–512.

Sun, C., Adamopoulos, P., Ghose, A. and Luo, X. (2022), 'Predicting stages in omnichannel path to purchase: A deep learning model', *Information Systems Research* **33**(2), 429–445.

Tang, J. (2024), 'Dcrmta: Unbiased causal representation for multi-touch attribution', *arXiv preprint arXiv:2401.08875* .

Tao, J., Chen, Q., Snyder Jr, J. W., Kumar, A. S., Meisami, A. and Xue, L. (2025), 'A graphical point process framework for understanding removal effects in multi-touch attribution', *Management Science* **71**(9), 7312–7332.

Thygesen, A. (2022), 'Reach Customers as They Browse and Buy'. [Accessed 3 November 2024].
URL: *https://www.thinkwithgoogle.com/intl/en-apac/consumer-insights/consumer-journey/reach-customers-on-google/*

Tsamardinos, I., Brown, L. E. and Aliferis, C. F. (2006), 'The max-min hill-climbing Bayesian network structure learning algorithm', *Machine Learning* **65**(1), 31–78.

Tueanrat, Y., Papagiannidis, S. and Alamanos, E. (2021), 'Going on a journey: A review of the customer journey literature', *Journal of Business Research* **125**, 336–353.

Twenge, J., Martin, G. and Spitzberg, B. (2019), 'Trends in US adolescents' media use, 1976–2016: The rise of digital media, the decline of TV, and the (near) demise of print', *Psychology of Popular Media Culture* **8**(4), 329.

Urbonavičius, S. and Dikčius, V. (2008), 'Importance of marketing activities in a company: the management perspective', *Ekonomika* **83**, 41–50.

Verhoef, P., Kannan, P. and Inman, J. (2015), 'From multi-channel retailing to omni-channel retailing: Introduction to the special issue on multi-channel retailing', *Journal of Retailing* **91**, 174–181.

Vollrath, M. and Villegas, S. (2021), 'Avoiding digital marketing analytics myopia: Revisiting the customer decision journey as a strategic marketing framework', *Journal of Marketing Analytics* **10**(2), 106–113.

Voorhees, C., Fombelle, P., Gregoire, Y., Bone, S., Gustafsson, A., Sousa, R. and Walkowiak, T. (2017), 'Service encounters, experiences and the customer journey: Defining the field and a call to expand our lens', *Journal of Business Research* **79**, 269–280.

Wernerfelt, N., Tuchman, A., Shapiro, B. T. and Moakler, R. (2025), 'Estimating the value of offsite tracking data to advertisers: Evidence from Meta', *Marketing Science* **44**(2), 268–286.

Wirtz, J., Tuzovic, S. and G. Kuppelwieser, V. (2014), 'The role of marketing in today's enterprises', *Journal of Service Management* **25**(2), 171–194.

Wymbs, C. (2011), 'Digital marketing: The time for a new "academic major" has arrived', *Journal of Marketing Education* **33**(1), 93–106.

Xie, X., Du, F. and Wu, Y. (2020), 'A visual analytics approach for exploratory causal analysis: Exploration, validation, and applications', *IEEE Transactions on Visualization and Computer Graphics* **27**(2), 1448–1458.

Xu, L., Duan, J. and Whinston, A. (2014), 'Path to purchase: A mutually exciting point process model for online advertising and conversion', *Management Science* **60**(6), 1392–1412.

Xu, Z., Vail, C., Kohli, A. and Tajdini, S. (2021), 'Understanding changes in a brand's core positioning and customer engagement: A sentiment analysis of a brand-owned Facebook site', *Journal of Marketing Analytics* **9**(1), 3–16.

Yadagiri, M. M., Saini, S. K. and Sinha, R. (2015), A non-parametric approach to the multi-channel attribution problem, *in* 'International Conference on Web Information Systems Engineering', Springer, pp. 338–352.

Yang, D., Dyer, K. and Wang, S. (2020), 'Interpretable deep learning model for online multi-touch attribution', *CoRR* .

Yao, D., Gong, C., Zhang, L., Chen, S. and Bi, J. (2022), CausalMTA: Eliminating the user confounding bias for causal multi-touch attribution, *in* 'Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining', pp. 4342–4352.

Zaeem, R. N. and Barber, K. S. (2020), 'The effect of the GDPR on privacy policies: Recent progress and future promise', *ACM Transactions on Management Information Systems* **12**(1), 1–20.

Zamil, A., Nusairat, N., Vasista, T., Shammot, M. and Areiqat, A. (2021), 'Prediction of sales based on an effective advertising media sale data: A Python implementation approach', *Academy of Strategic Management Journal* **20**(25).

Zenetti, G., Bijmolt, T. H., Leeflang, P. S. and Klapper, D. (2014), 'Search engine advertising effectiveness in a multimedia campaign', *International Journal of Electronic Commerce* **18**(3), 7–38.

Zhang, Y., Wei, Y. and Ren, J. (2014), Multi-touch attribution in online advertising with survival theory, *in* '2014 IEEE International Conference on Data Mining', IEEE, pp. 687–696.

Zhao, K., Hua, J., Yan, L., Zhang, Q., Xu, H. and Yang, C. (2019), A unified framework for marketing budget allocation, *in* 'Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining', pp. 1820–1830.

Zhao, K., Mahboobi, S. H. and Bagheri, S. R. (2018), 'Shapley value methods for attribution modeling in online advertising', *arXiv preprint arXiv:1804.05327* .

Zhao, K., Mahboobi, S. H. and Bagheri, S. R. (2019), 'Revenue-based attribution modeling for online advertising', *International Journal of Market Research* **61**(2), 195–209.

Zhou, H., Li, S., Jiang, G., Zheng, J. and Wang, D. (2023), Direct heterogeneous causal learning for resource allocation problems in marketing, *in* 'Proceedings of the AAAI Conference on Artificial Intelligence', Vol. 37, pp. 5446–5454.

Zia, M. and Rao, R. (2019), 'Search advertising: Budget allocation across search engines', *Marketing Science* **38**, 1023–1037.