UHASSELT
KNOWLEDGE IN ACTION

Faculteit Bedrijfseconomische Wetenschappen

master handelsingenieur in de beleidsinformatica

Masterthesis

Evaluating the Impact of Environmental Award Criteria in EU Construction Public Procurement: Assessing the Alignment of GPP with the Principles of Non-Discrimination and Competition

Lucas Ricour
Scriptie ingediend tot het behalen van de graad van master handelsingenieur in de beleidsinformatica

PROMOTOR :
Prof. dr. Sebastien LIZIN

COPROMOTOR :
Prof. dr. Stephan BRUNS

BEGELEIDER :
De heer Haitham ABU GHAIDA

UHASSELT
KNOWLEDGE IN ACTION

www.uhasselt.be
Universiteit Hasselt
Campus Hasselt:
Martelarenlaan 42 | 3500 Hasselt
Campus Diepenbeek:
Agoralaan Gebouw D | 3590 Diepenbeek

2024
2025

Faculteit Bedrijfseconomische Wetenschappen

master handelsingenieur in de beleidsinformatica

Masterthesis

Evaluating the Impact of Environmental Award Criteria in EU Construction Public Procurement: Assessing the Alignment of GPP with the Principles of Non-Discrimination and Competition

Lucas Ricour

Scriptie ingediend tot het behalen van de graad van master handelsingenieur in de beleidsinformatica

PROMOTOR :

Prof. dr. Sebastien LIZIN

COPROMOTOR :

Prof. dr. Stephan BRUNS

BEGELEIDER :

De heer Haitham ABU GHAIDA

Evaluating the Impact of Environmental Award Criteria in EU Construction Public Procurement: Assessing the Alignment of GPP with the Principles of Non-Discrimination and Competition

Lucas Ricour[2054759]

Hasselt University - Masterthesis BI (3528) - Academic Year 2024-2025

Keywords: Directive 2014/24/EU · Construction · Public Procurement · Green Public Procurement · Sustainable Public Procurement · Environmental Criteria · Difference-in-Differences · Tender Electronic Daily.

Preface

The motivation for this research stemmed from my interest in the intersection of data analysis and sustainability. This thesis offered a unique opportunity to combine both areas by empirically exploring how environmental objectives are incorporated into public procurement. Throughout the research process, I gained valuable insights into working with large and complex datasets, addressing data quality issues, and designing robust empirical analyses, all of which will be highly relevant for my future career. My interest in sustainability has been a recurring theme throughout my academic journey, culminating in a particularly enriching international experience in Indonesia through the EcoGreen project. These experiences broadened my understanding of global sustainability challenges and strengthened my resolve to pursue socially relevant research. Since the start of this thesis in September, I have been challenged, put under pressure, and pushed beyond my comfort zone. In navigating these challenges, I learned to adapt, persist, and problem-solve in the face of uncertainty. It also required me to apply and refine the skills I had developed throughout my academic journey, bringing together everything I have learned over the past years. I would like to express my sincere gratitude to my supervisor, Prof. Dr. Sebastien Lizin, my co-supervisor, Prof. Dr. Stephan Bruns, and my advisor, Mr. Haitham Abu Ghaïda, for their valuable guidance and support throughout this project. I am also deeply thankful to my parents, Caterina and Dirk, my brother Alexander, and my friends for their continuous encouragement and support.

Summary

This thesis analyzes the effect of the 2014 Public Procurement Directive on construction public procurement in the European Union (EU), with a distinct focus on the impact of the inclusion of environmental (green) award criteria in construction works call for tenders on the number of bidders (used as a proxy for competition) and project costs, which are approximated using the tentative contract values due to the unavailability of actual cost data. The thesis seeks to contribute to the ongoing academic and policy debate by providing empirical evidence on the extent of misalignment between the observable impacts of utilizing green award criteria and the fundamental principles of the EU internal market (competition, transparency, and equal treatment).

To estimate a causal relationship between green criteria utilization and procurement outcomes, the research employs data from the EU Tenders Electronic Daily (TED) database from 2012 to 2023 to

operationalize a one-way fixed effects Difference-in-Differences (DiD) specification in an aggregated two-period model at the Contracting Authorities or Entities (CAE) level. The dataset is filtered only to include 'Works' contracts above the EU value threshold. Observations from Liechtenstein (LI), the United Kingdom (UK), Greece (GR), Switzerland (CH), and North Macedonia (MK) are excluded due to coverage limitations. The methodology's robustness is assured by clustering standard errors at the CAE level to account for potential correlations within each observed entity while controlling for unobserved heterogeneity within individual CAEs using fixed effects.

The results indicate that using green criteria significantly (at the 5% level) reduces the number of tenders received by 15.1% to as much as 23.7%, suggesting that environmental criteria influence market participation. This thesis cannot assess whether the observed effect on market participation fades over time, as the applied two-period aggregation does not allow for dynamic or time-varying impact analysis. The study identified and acknowledged that the estimated decline cannot be directly attributed to the pure inclusion of environmental criteria, as these criteria do not explicitly or legally prohibit suppliers from submitting bids. The research further shows a moderate rise in project cost if CAEs implement environmental criteria after the announcement of the 2014 directive, with significant (at the 10% and 5% level respectively) increases ranging from 14.6% to 17.5%. However, statistical significance depends on the definition of the applied treatment. Non-significant increases of 5.2% to as much as 8.1% can be observed in less strict settings, implying that the effect relies on the application intensity of green criteria or the inherent price difference between green and non-green contracts.

While the research design distinguishes between consistent and partial use of environmental award criteria, it does not directly compare fully green to mixed adopters. As a result, no conclusions can be drawn about the role of adoption intensity in project costs. In contrast, the effect on bidding competitiveness appears more robust, with a consistent decline in bids observed across all models and configurations. The results, while informative, are sample-specific, given the within-unit identification of treatment effects and heterogeneity in the procurement practices across CAEs. Generalization of any causal statement must, therefore, be done cautiously. Even though introducing environmental criteria is expected to foster sustainability in the Single Market, the results show that it unintentionally creates (soft) barriers to competition, disrupting the core principles of the EU internal market. This thesis opposes the EU's beliefs about the impact and potential of Green Public Procurement (GPP) as a strategic policy tool, proving that the balance between fostering sustainability and offering fair competition has not been achieved.

Based on the insights derived from the existing literature and empirical analysis, this thesis advises that the application of the 2014 Directive needs to be reviewed further, calling for a reevaluation of its implementation to better align sustainability objectives with the principles of free market competition.

Future research should investigate the underlying mechanisms driving the observed decline in bidding competitiveness, such as suppliers' perceived compliance costs, strategic self-selection, or fears of biased procedures. Additionally, there is a need to explore whether reduced competition in GPP represents an unintended consequence or a deliberate trade-off in pursuit of greater environmental gains. This includes evaluating whether the environmental benefits of green procurement sufficiently outweigh the additional costs to public authorities. As regulatory frameworks evolve, notably with the introduction of mandatory sustainability requirements under the Ecodesign for Sustainable Products Regulation (ESPR), further empirical research is needed to evaluate how these measures influence market access and supplier diversity. Additionally, working towards more standardized and detailed procurement data would significantly improve the ability to analyze the impact of green criteria on both competition and pricing outcomes.

Table of Contents

1 Introduction

In recent decades, the European Union (EU) has begun actively engaging and supporting partners and Member States in taking action against the global threat of climate change [19]. The European Commission (EC) considers climate change as the greatest challenge of our time, but also an opportunity for change [22]. Ambitious initiatives such as the European Green Deal, aiming to transform the EU economy into a modern, resource-efficient, and competitive one, are ensuring actions to achieve climate-neutrality by 2050 are taken on multiple fronts [25]. The European Green Deal strives for carbon neutrality, economic growth decoupled from resource consumption, and to leave no person or place behind [25]. The Green Deal serves as a blueprint, harmonizing existing directives with newly brought forth ones, legally binding climate targets, and comprehensive proposals for changes covering all sectors of the economy.

One such existing directive is the Public Procurement Directive 2014/24/EU[1], which focuses on the public purchase of services, goods, and works by public entities (i.e., local & (inter)national governments and authorities, public healthcare providers, public educational institutes, etc.), a process better known as Public Procurement (PP) [6][27][42][67]. Essentially, the Commission's objectives regarding this directive were twofold: Improving the efficiency of procedures while enabling the more strategic use of public procurement to advance environmental, social, and industrial or innovation-related policy objectives. [13]. The directive enabled the strategic usage of procurement by allowing contracts to be awarded not solely on price but also on non-price criteria, such as environmental criteria [53].

Environmental criteria are a way to identify green products and green suppliers [69]. In Yu et al. (2020) [69], they are described as being a broad combination of environmental aspects and procurement criteria that are essential requirements defined by the demanding party that must be met by the possible supplier(s) for the whole duration of the contract. In Fuentes-Bargues et al. (2019) [32], a list of identified environmental criteria is presented. A criterion can, for example, be broad, such as reduced emissions or resource consumption, or more specific, like required certified environmental accreditation or obliged use of particular materials and techniques [32].

The PP directive enables and acknowledges the importance and inclusion of environmental criteria at different phases of the tendering process [46]. Wang et al. (2020) [68] propose a simplified structure of the procurement cycle, distinguishing three high-level phases: planning, purchasing, and contract execution. Although criteria can, to some extent, be applied in every phase, they are most commonly applied during the purchasing phase in the form of award criteria, where they are used to evaluate the environmental considerations of tenders. Testa et al. (2016) [63] based their analysis on the five procurement components where environmental criteria can be applied: Subject matter, selection criteria, technical specifications, award criteria, and contract performance clauses.

Carreras (2023) [6] considers PP one of the main instruments for transferring public funds to the private market. This transfer makes up a noteworthy slice of the economy, globally between 10-15% of the Gross Domestic Product (GDP) of developed countries [6][32][33][42][67]. Studies show that this figure lies between 14-16% for Member States of the EU [42][49][59][62][63].

The substantial purchasing power held by the demand side of PP transactions makes the process a strategic tool for the EU to aid and promote its sustainability goals [6][49]. Kumar (2022) [42] and Nemec et al. (2022) [49] state that PP has the potential to drastically reshape how public funds are expended by changing production- and consumption behavior, addressing market failures, increasing procurement process efficiency, and improving economic growth, while simultaneously incentivizing for being environmentally friendly [42][49].

[1] Builds upon Directive 2004/18/EC

This tendering approach is called Green Public Procurement (GPP), which falls under the broader concept of Sustainable Public Procurement (SPP) [6][32]. Within PP, efficiency is comprised of more than just cost-effectiveness; it encompasses time and labor efficiency, as well as operational effectiveness, the latter referring to how well procurement processes are executed to meet their intended goals (quality, compliance, and timely delivery) [68]. Factors affecting efficiency within PP include competition, transparency, the number of bidders, decentralization, and evaluation methods [68].

SPP is described as a process by which public authorities seek to achieve the appropriate balance between the three pillars of sustainable development (economic, social, and environmental) when procuring goods, services, or works at all stages of the project [15]. These three pillars align with the concept of the Triple Bottom Line, referred to as the three Ps: People, Planet, and Profit [15]. GPP is less broad and is primarily focused on environmental sustainability [6][33].

The EC delineates GPP as a process whereby public authorities seek to procure goods, services, and works with a reduced environmental impact throughout their life cycle when compared to goods, services, and works with the same primary function that would otherwise be procured, a definition generally accepted by researchers and academics[2] [26].

Despite ongoing efforts to promote GPP, it remains underutilized, even with the use and support of legislative instruments [48][55][59]. Sönnichsen & Clement (2020) [62] described the adoption and transformation as nascent. Even though by 2010, most EU Member States had developed GPP action plans, GPP contracts seldom accounted for more than 40% of public procurement expenditure [48][55].

Prier et al. (2018) [55] blames the lack of uptake on legal challenges and officials' reluctance towards risk, with other studies adding officials' lack of expertise regarding GPP contracts [48][63]. Adding environmental criteria to PP contracts also raises the complexity [48][62]. Another highlighted challenge hindering adoption is the lack of well-defined environmental criteria [33][32][48].

Carreras (2023) [6] suggests that evidence about cost prices of GPP contracts is scarce and mixed, adding that due to the heterogeneous nature of GPP, it is tough to extrapolate results, with actual direct costs being so subjective to specific contexts. This heterogeneity refers to the variation in the usage of environmental criteria across sectors and regions and inconsistencies in how GPP is defined, mandated, and implemented.

Van Assche et al. (2024) [67] fear that including environmental criteria will create a public market environment that excludes certain companies and favors others, affecting competitiveness and bidding behavior. In Yu et al. (2020) [69], the researchers concluded that despite years of trials and changes, it remains unclear what exactly promotes and hinders the utilization of GPP.

Given ongoing debates about GPP's dual role as a sustainability instrument and a market mechanism within the EU. This study aims to contribute to a better understanding of the real-world implications. GPP has increasingly been positioned as a strategic policy tool to align public spending with environmental objectives [44][49]. However, GPP has been hypothesized to introduce entry barriers that may favor certain suppliers [67]. Hampering its compatibility with the principles of non-discrimination and free competition. These principles require procurement processes to ensure equal access, transparency, and competition. If GPP unintentionally undermines these conditions, it may reduce supplier diversity and challenge the coherence of the internal market.

GPP is often described in policy discussions as a tool for driving long-term change and promoting sustainability. However, studies by Lundberg et al. (2015) [45] and Lundberg and Marklund (2018) [44] show that there is little evidence that GPP leads to such changes in practice. Their research

[2] Definition adopted by the following sources: [33][32][42][43][48][49][55][62][63][69]

suggests that the way GPP is currently designed has had only a small impact on supplier participation, raising doubts about its effectiveness as an environmental policy instrument [44][45].

Against this backdrop, this study seeks to estimate the actual effects of including environmental criteria in procurement. It does so by analyzing the tendering behavior of Contracting Authorities or Entities (CAEs) and comparing outcomes of calls for tenders, specifically bid competitiveness and project cost, that include environmental considerations with those that do not.

The focus lies on the construction sector due to its importance in both sustainability and public procurement. It is one of the EU's most procurement-intensive sectors and carries significant environmental impacts, such as emissions, resource utilization, and waste generation [32][33]. Moreover, it is well represented in the Tenders Electronic Daily (TED) database, providing a robust empirical basis for the study. TED is the EU's official online platform for public procurement, publishing tenders that exceed specific value thresholds[6][55][59]. It promotes transparency and enables tracking of contracts, including whether environmental criteria are applied [55][59].

The remainder of this paper is structured as follows: Section 2 gives a structured overview of the research questions. Section 3 offers a more in-depth description of several key concepts essential for understanding the paper, including the role and challenges of Public Procurement (PP) in the EU, the distinctions between Sustainable, Green, and Circular Public Procurement, the phases of Green Public Procurement (GPP), its potential as a policy tool, challenges in implementation, methods for identifying green contracts and measuring competition, and the significance of GPP in the construction sector. In Section 4, the methodology used in this research will be outlined. Detailing the data collection, cleaning, transformation, and other preprocessing steps. Afterward, the research design, assumptions, and analysis methods used to assess the effects of environmental criteria on project prices and bid competition in the construction sector are discussed. Section 5 outlines the results of the quantitative analysis, focusing on how environmental criteria impact project costs and bid competition. Finally, in Section 6, the research contribution to academic knowledge and practical implications of the discovered findings are underlined and discussed, In addition recommendations for future research are made, while addressing challenges encountered during and limitations of the study.

2 Research Questions

While conventional procurement frequently emphasizes the lowest price, the EU's latest policy shifts foster the incorporation of additional criteria to serve as a breeding ground for all parties involved to engage in innovation and sustainability in a controlled environment to achieve broader public value. Such criteria involve quality, innovation, sustainability, and social value. While social value, such as the reduction of environmental harm, can be an important justification for GPP, this thesis does not aim to measure such socially valuable outcomes. Instead, it explicitly investigates the role of environmental award criteria in PP within the EU construction sector and how these criteria impact procurement outcomes, specifically, bidding competitiveness and project cost. A set of three research questions is formulated to guide the research and analysis of the dataset. An overview of the questions is given below:

Scope: All research questions focus on EU public procurement 'Works' contracts in the construction sector, above the EU threshold, between 2012 and 2023. The analysis is limited to all EU countries except Liechtenstein (LI), the United Kingdom (UK), Greece (GR), Switzerland (CH), and North Macedonia (MK), which were excluded due to data limitations.

- **RQ1:** *To what extent are environmental award criteria used in construction public procurement contracts?*

- *Prerequisite: How can award criteria be reliably categorized into price vs. environmental, using available data?*

– **RQ2:** *To what extent does the integration of environmental criteria in construction public procurement contracts influence bidding competitiveness, either positively or negatively?*
  - *Prerequisite: How can bidding competitiveness be quantified using available data?*

– **RQ3:** *To what extent is the integration of environmental criteria into construction public procurement contracts responsible for the increase/decrease in project cost?*
  - *Prerequisite: How are project costs quantified?*

RQ2 and **RQ3** are the main research questions of this study. **RQ1** is specifically formulated to aid in answering **RQ2** and **RQ3** successfully. **RQ1** is a prerequisite to grasping the baseline utilization, necessary scope, and possible variability. Answering **RQ1** could reveal low inclusion of environmental criteria in certain regions. Such context can be a reason to halt further analysis or be crucial for interpreting insubstantial results on bidding competitiveness (**RQ2**) or project costs (**RQ3**) in those regions. Furthermore, if criteria are amply applied, this would justify further analysis.

3 Background

This section gives an overview of the context and several key concepts for understanding the remainder of the paper. First, Section 3.1 introduces Public Procurement within the European Union, outlining its key role in achieving broader policy objectives, discussing how this role evolved over time, and highlighting some tensions, pressures, and challenges it faces. Section 3.2 explores the distinctions and interconnectedness between Sustainable, Green, and Circular Public Procurement to reduce ambiguity and provide explicit scopes of the concepts. Section 3.3 describes the key phases of the Green Public Procurement process, detailing how environmental considerations are integrated at various stages, focusing mainly on the integration within the construction sector throughout a running example. Section 3.4 discusses the significant promise of GPP as a powerful policy tool for achieving environmental objectives whilst considering concerns raised in the literature. Section 3.5 dives deeper into the challenges that hinder the successful translation of the strong ambitions surrounding GPP. Crucial to this research is the identification of "green" contracts. To provide a basis for answering this prerequisite, Section 3.6 discusses how previous studies handled this. Of similar importance is the way competitiveness can be quantified; therefore, Section 3.7 explores different approaches to measuring competition in public procurement, offering insights into relevant analytical methods. Finally, Section 3.8 delves into the role of GPP within the construction sector, emphasizing the importance of this industry and the detrimental strain it has on the environment. Additionally, the potential impact GPP can have on the adverse side effects of construction is presented.

3.1 Public Procurement in the European Union

Public Procurement (PP), defined as the public purchase of services, goods, and works by public entities [27], is a cornerstone of the European economy and has become an important policy tool for enhanced supranational integration [53]. With an estimated 14% of the EU's GDP being spent through PP and over 250.000 public authority participants, it is considered a fundamental element of the investment ecosystem. Within the EU, procurement plays a crucial (geo)political role in achieving broader policy objectives [15][53].

The EC has established a minimal set of rules and guidelines to create efficient, harmonized procurement markets in the EU in order to level the playing field for businesses [27][53]. One such regulatory mechanism is threshold values, which determine whether a procurement contract falls under EU or national regulations. These thresholds vary depending on the type of contract, industry and contracting authority. The primary limits for construction projects are €143 000 for most types of supplies and services procured by central government authorities; for construction works contracts, the threshold is €5 538 000. Contracts exceeding these levels are obliged to comply with EU-wide procurement procedures. While these procedures can differ in form (open, restricted, negotiated), this thesis cannot examine procedural variation due to the aggregation of data at a higher level, this limitation is further addressed in Section 6. National PP rules apply for lower-value tenders, but the EU principles of transparency and equal treatment should still be respected.

Pircher (2022) [53] identified the EU's six strategic priorities as being increased uptake of green and social criteria, improved professionalization of public procurers, improved transparency to counteract corruption, ease of access to global markets, joint procurement, and magnifying the digital transformation within procurement.

With the sustainability of actions being a core value of the EU, it aims to maintain the Single Market principles of free movement of goods, equal treatment, and transparency to ensure fair competition, non-discrimination, and opportunities across Member States [51]. By integrating sustainability into these principles, the EU seeks to balance economic growth with environmental responsibility [42][49].

With challenges and objectives such as achieving Green House Gas (GHG) emissions reductions to comply with the targets set in the Paris Agreement and European Green Deal, the EU tries to harness PP to produce spillover effects, especially in large infrastructure projects, which require significant amounts of material resources [60]. However, little evidence exists on whether these policies produce the intended spillover effects, such as increased private adoption of green standards, development of local expertise, improved supply chains, and greater visibility and legitimacy for sustainable practices [60].

The evolution of EU PP directives reflects the EU's desired objective to align PP with sustainability goals. The 2004 directives allowed procurers to award tenders based on criteria other than just the price [52]. This concept was further refined and defined in the 2014 directives, which also introduced environmental considerations concerning award criteria, technical aspects, and construction conditions [3][14]. The 2014 directive aimed to both reduce the environmental impact of public expenditures and stimulate innovation and adoption of greener alternatives [13][55].

Historically, EU procurement policy had an endogenous source of change, resulting from policy learning driven by case law and challenges in application [53]. However, the economic crisis of 2008 acted as a catalyst for significant reforms [53]. With the 2014 directives on public procurement, the EC aimed to harmonize procurement at a European level and sought to overcome domestic constraints, referring to national protectionism, and the limited capacity or willingness to align with EU goals [53].

However, PP is a policy field characterized by contradictory pressures and tensions. While the EC strives for greater liberalization and competition to widen their procurement markets, Member States often refrain from adhering in order to protect national economies [53]. Additionally, the EU's continuous efforts to motivate Member States to utilize PP for other political objectives, such as stimulating economic recovery or promoting environmental goals, have had varying outcomes due to Member States' differences in their willingness and capacity to align with these directives [53].

Pircher (2022) [53] argues that whilst PP has seen increased European integration, it has not led to full market integration in practice because EU PP operates as a shared competence. This means that it is subject to the principle of subsidiarity—the EU sets rules to ensure the functioning of the internal market, but Member States can implement these rules at their own discretion—as shown in Figure 1 [53]. Unlike regulations, which are directly enforceable across all Member States, directives set objectives that must be achieved by each country [50]. With directives, Member States can design their own frameworks and action plans. This flexibility introduces variations in how directives are applied and when they are applied [50]. The EU can take legal action if directives are not properly implemented, but this process is lengthy and can be resisted by Member States. For this reason, directives can appear to be improperly applied or seemingly ignored [50].

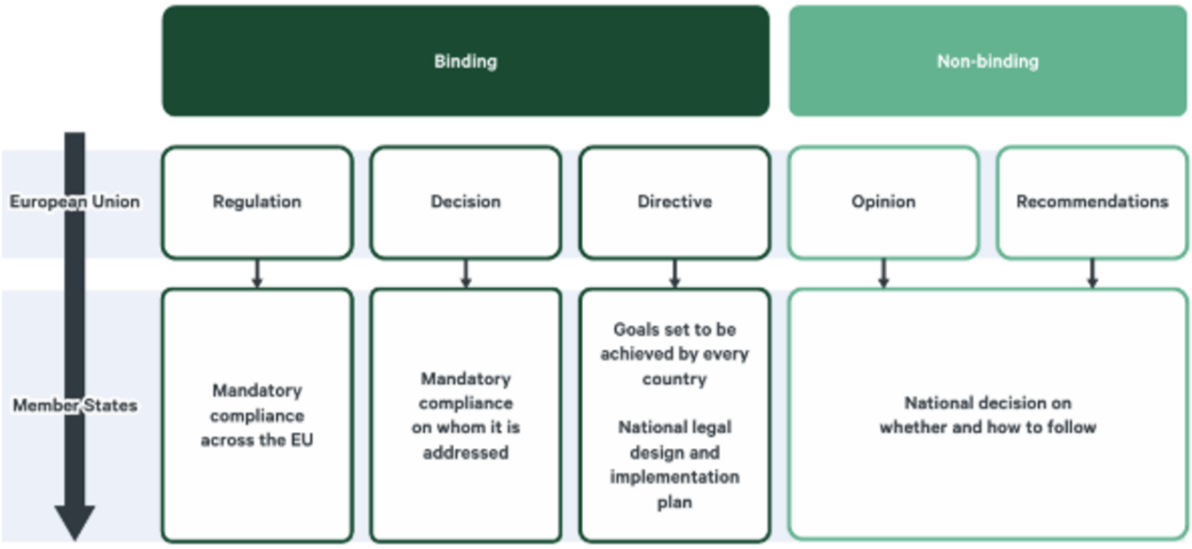


Fig. 1: EU legislation across governance levels. Reprinted from *Green public procurement: a key to decarbonizing construction and road transport in the EU*, by Nilsson Lewis et al., 2023. [50]

3.2 Comparing Sustainable, Green, and Circular Public Procurement

This section explores the distinctions between Sustainable, Green, and Circular Public Procurement, providing a foundational understanding of these procurement concepts in order to reduce the ambiguity surrounding the distinctions and interconnections between them.

Figure 2 shows the interconnectedness of the three types of procurement. SPP is considered to be an overarching concept, encompassing both Green and Circular Public Procurement, with CPP being a subset of GPP [36]. Ahmed et al. (2024) [3] and Nilsson Lewis et al. (2023) [50] highlight that terminology tends to be used interchangeably. This research finds such usage to be inconsistent. All three types are seen as separate concepts based on their scope and focus areas, albeit slightly overlapping.

Definitions of SPP vary in the literature [15]. In this research, the definition presented by the EC has been adopted. SPP is seen as a process by which public authorities seek to achieve the appropriate balance between the three pillars of sustainable development—economic, social, and environmental—when procuring goods, services, or works at all stages of the project [15]. SPP policies are developed to urge suppliers to aid in achieving societal objectives, ranging from supporting SMEs and promoting fair labor to reducing emissions and fostering innovation [36].

GPP directly prioritizes a reduction in environmental impact when considering the entire life-cycle of the procured products [3]. With CPP, the scope is slightly narrower, focusing mainly on material and energy efficiency, as well as waste prevention and reduction [50].

The GPP process involves screening offers against mandatory environmental criteria. Only those meeting the ’green’ standards are deemed eligible offers [45]. The term environmental criteria encompasses a broad combination of environmental aspects and procurement criteria that are essential requirements defined by the demanding party that must be met by the possible supplier(s) for the whole duration of the contract [69]. GPP balances environmental protection objectives with the integrity of the internal market [51]. This balance can become problematic when green criteria are applied in ways that restrict competition or reduce transparency, thus risking violating the core principles of the internal market, namely non-discrimination, equal treatment, and transparency.

The Circular Economy is an economic model in which planning, resourcing, procurement, production, and reprocessing are designed and managed to maximize ecosystem functionality and human well-being [64]. This strict definition implies no negative environmental effects [64]. Stemming from this CPP is defined by the EC as an approach to Green Public Procurement that pays special attention to the purchase of works, goods, or services that seek to contribute to the closed energy and material loops within supply chains whilst minimizing and in the best case avoiding, negative environmental impacts and waste creation across the whole life-cycle [64].

In Tátrai and Diófási-Kovács (2021) [64], it is stated that CPP acknowledges the important roles private as well as public authorities play in facilitating the transition towards a circular economy. The EU released its first Circular Economy Package in 2015, aiming to promote recycling, prevent material loss, create jobs, foster economic growth, advance zero-waste initiatives through eco-design and industrial symbiosis, and reduce greenhouse gas emissions and environmental impacts.

This section highlights SPP as a broad framework, GPP as focused on environmental impacts across the lifecycle, and CPP as emphasizing closing loops within supply chains to increase resource efficiency and waste reduction, showcasing their distinct scopes within sustainable procurement.

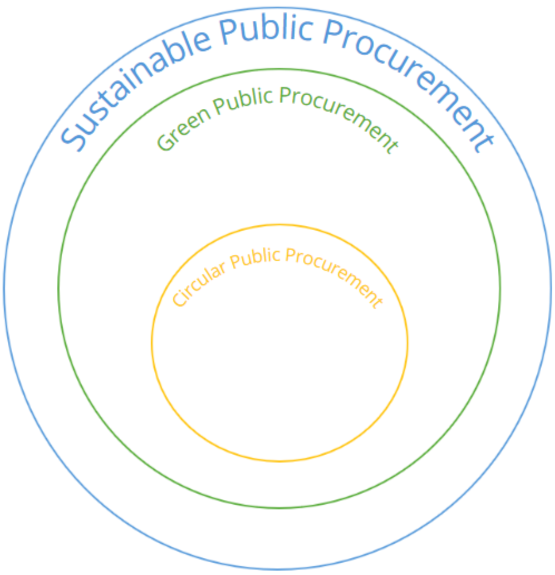


Fig. 2: Hierarchical relationship between SPP, GPP, and CPP, illustrating their interconnected scope and focus areas.

3.3 Green Public Procurement Process Phases

This subsection delineates the key phases of the Green Public Procurement process, detailing how environmental considerations are integrated at various stages. With GPP in construction being the main focus of this research, particular attention is given to how these phases apply within the construction sector, illustrated through a running example.

In this study, we adopt the high-level structure proposed by Wang et al. (2020) [68], viewing the GPP process as consisting of three main phases: planning, purchasing, and contract execution.

Given that this thesis focuses on works contracts, it is noteworthy that the full tendering cycle for a commonly used procedure, the international open tender, typically spans a minimum of seven to eight months. [21], as illustrated in the procedural timeline in Appendix 7.16. For a detailed explanation of each procedural step, readers are referred to *Contract Procedures for EU External Action: A practical guide* [21], as the running example in this section focuses on the high-level phases only.

The **planning phase** starts with identifying and acknowledging a specific need [68], by performing preliminary market consultations and analyzing public demand [64]. Afterward, the needed documentation is prepared [68]. In this stage, the contracting authority will decide if green aspects will be implemented and stipulated in the contract notice [64]. If indeed this is the case, contracting authorities are obliged to set out these environmental requirements in the call for tenders [51]. The contract notice will encompass details, including auction rules, award criteria (green or other), selection criteria, grounds for supplier exclusion, and the technical specifications (the so-called subject matter of the contract) [3][14][45][64]. In Dodd et al. (2016) [11], the GPP criteria, as initially outlined in the Buying Green Handbook, are described as follows: Selection criteria focus on determining the ability of the bidder to perform the contract and adhere to the requirements. The technical specifications outline the scope and characteristics of the work, product, or service. Award criteria can encourage higher environmental performance and innovation. When a contracting authority uses the Most Economically Advantageous Tender (MEAT) approach as the contract award criterion, it shall specify the relative weighting of each criterion [51]. MEAT is a valuation and selection method allowing contracting authorities to award contracts based on characteristics other than just the purchase price [6]. When offers are eventually received, they will be screened, and the bidder whose total number of 'points' is highest will be awarded the contract [51]. In Parikka-Alhola and Nissinen (2012) [52], this awarding procedure was scrutinized. More specifically, the calculation method was questioned for not ensuring competitive bidding. Explaining that bidders who lack the resources or expertise to conduct detailed life-cycle assessments, or those from smaller companies and countries with limited access to environmental data are disadvantaged [52].

However, the prominent role assigned to MEAT in the 2014 EU directives was intended to address such concerns, as it emphasizes a balanced evaluation of tenders based on economic and qualitative factors. Theoretically, these environmental characteristics should not hurt bidding competitiveness, as they are meant to be non-discriminatory towards potential bidders [51]. This non-discriminatory nature is ensured by the obligation of contracting authorities to accept equivalent proofs of compliance, like Eco-labels, technical documentation, or certificates, which helps prevent the exclusion of capable suppliers based on overly specific or localized requirements [50]. There are also legal requirements that state that environmental criteria must be objective, measurable, and linked directly to the contract's subject matter [14][45]. For example, if a contract notice for a construction project requires that all employees commute to and from the construction site using electric vehicles, it would impose an unnecessary barrier to competition and lack a direct connection to the construction process. Unique to GPP is that during this phase, contracting authorities will consult external experts to aid in detailing the use of Eco-labels, formulating green criteria, and setting certification targets [3]. Ahmed et al. (2024) [3] address the significant role GPP plays in shaping tender documents, referencing the high number of choices that need to be made prior to sending out the call for tenders.

Running Example: While conducting a survey of its population, a municipality identified a demand for more leisure infrastructure. In response, it was decided to build a new sustainable community center that would serve as a multipurpose hub for celebrations, meetings, and sporting events. Several ways to reduce the building's environmental impact were identified during a preliminary market consultation. With the help of external experts, the feasibility of these options was discussed. Based on their advice, the following technical specifications were included in the contract notice: the installation of solar panels, an advanced rainwater collection system to provide part of the water supply, the usage of low-emission machinery, and the usage of sustainably sourced and certified building materials. The municipality will use the MEAT approach as an award criterion, considering other aspects than just the price. For example, bidders can receive extra credit or points if they propose innovative or more energy-efficient solutions that decrease the project's life-cycle costs. To ensure only qualified bidders participate, the municipality applies selection criteria. In this case, the firms must present a detailed timeline of the building process and a list of certified and sustainable suppliers. If bidders fail to meet these requirements or have incurred environmental violations in the past, they can be excluded from the bidding process.

The **purchasing phase** can be divided into two stages: the tender stage and the post-tender stage [68]. In the tender stage, the aforementioned documents are published, also known as announcing the call for tenders. This marks the initiation of the procurement auction, which can include one or several lots. A supplier can bid on a subset or all contracts within a single procurement call [14][45]. The bidding can co-occur but needs to be independent across contracts [14]. At this stage, the contracting authority must also address questions from market players regarding unclear aspects of the contract notice [64]. Tenders can be submitted until the predetermined deadline, which marks the end of the tender stage and the start of the post-tender stage. During the post-tender stage, the public body's main job is organizing bid opening meetings, evaluating submissions, and selecting the winning bid [68]. As previously mentioned, the most fundamental form of tender evaluation consists of checking compliance with the exclusion ground and selection criteria, as well as evaluating the award criteria [64]. This inquiry includes a rigorous analysis of submitted prices, focusing especially on abnormally low bids. Another best practice is to request assessments of the potential environmental impact and verify and substantiate claims made in the tender regarding (green) credentials and pollution reduction methods. This is all done to ensure the validity of submissions [64]. Although the EU procurement process is regulated by law, it offers contracting authorities significant freedom in crafting their bid evaluation processes as long as they comply with general principles such as equal treatment, transparency, and proportionality [14][45]. The latter essentially states that requirements and criteria must be reasonable and not more demanding than needed to fulfill the contract [30].

Running example: The municipality posts its call for tenders to construct the sustainable community center. The tender is divided into two lots. One concerns the construction of the main building structure, and the other concerns the installation of the solar panels and rainwater collection system. The call for tenders is posted on TED, and tenders can be submitted over a six-month period. During this period, several bidders submit questions for clarification, which are promptly addressed and published to ensure fairness. When the deadline passed, the municipality received two official bids for the first lot and three for the second. All submissions are reviewed during a bid-opening meeting. Firstly, all tenders are checked for compliance with the exclusion grounds and selection criteria. For each lot, a supplier is disqualified due to providing insufficiently elaborated project timelines. Only one supplier remains for the construction of the main building, but before awarding the contract, the municipality needs to ensure the submission is valid. After consulting the provided documents and assessments from an external expert, the supplier was notified of the acceptance of their bid, and for this contract, the contract execution phase begins. For the other lot, however, two potential suppliers remained. In this case, the municipality will first evaluate based on the award criteria. Supplier X proposes an innovative sustainable energy plan, promising massive long-term savings, which scores highly with the procurers. Supplier Y attached its portfolio of completed sustainable projects and references, highlighting its established track record. Both offers were then analyzed for their validity. Ultimately, Supplier X's bid was deemed abnormally low and therefore disqualified. As a result, Supplier Y was awarded the contract.

The **contract execution phase** commences with the signing of the contract and consists mainly of monitoring activities [68][3]. Tátrai and Diófási-Kovács (2021) [64] propose a two-stage phase. Firstly, in the execution stage, the contract's implementation is verified. This includes checking subcontractors and ensuring that any significant changes to commitments do not compromise or devalue the environmental objectives originally agreed upon [64]. The second stage, monitoring, initiates after project completion and evaluates the tender's success in meeting its demands. This involves assessing the resources expended during the procurement process, such as time and money, and calculating long-term sustainability outcomes like pollution reduction or water usage [64]. The monitoring stage extends beyond project completion. Synthesizing lessons learned and identifying best practices to inform and improve future procurement procedures.

Running example: After signing the contract, the municipality assigns a site manager to visit the construction site to ensure compliance with commitments regularly. On one visit, several infractions were noted. The contractor hired a subcontractor who began construction using uncertified materials and outdated machinery, contrary to the agreed terms. These issues were promptly raised with the contractor and resolved appropriately. No further issues arose during the project, and the building was completed. Post-completion, the municipality continuously evaluates outcomes by verifying the water savings from the rainwater system and the functionality and capacity of the solar panels and documenting lessons learned to enhance future procurement processes.

3.4 Green Public Procurement as a Powerful Tool for Policymakers

In a recent statement, Elina Bardram, Mission Manager for the EU Mission on Adaptation to Climate Change, highlights that 'environmental protection is not just about reducing emissions; it's about building resilience and preparing for the impacts we're already facing' [10]. Her statement accentuates the need for a holistic approach to environmental policy.

Within the EU, environmental protection at the macro and micro levels has gained importance among policymakers at an accelerating pace [48][49]. The Green Paper on Integrated Product Policy (IPP) reiterates the leading role public authorities must adopt to positively transform consumption patterns towards greener alternatives [51]. Increasing environmental distress concerning EU actions and policies has propelled the importance of green aspects within PP [51]. GPP has become an important instrument as the consensus that adding environmental regulations and criteria to PP can be effective measures against environmental challenges grows [6][62]. The Europe 2020 strategy substantiated these beliefs by exclaiming PP as a key market instrument to achieve smart, sustainable, and inclusive growth using public funds in the most efficient manner [13][46]. Ahmed et al. (2024) [3] even claim EU policymakers consider GPP crucial for achieving environmental policy objectives, resulting in a gradual integration of GPP to limit the negative environmental impact of their procurement activities [56].

In Nemec et al. (2022) [49], it was noted that the perception of PP shifted towards an increasingly more strategic one, adding to its original purpose of fulfilling public demands. In agreement with [38][49][69], Van Assche et al. (2024) [67] explain that PP is a demand-led policy tool primarily designed so that governments can act as large discretionary buyers in an attempt to steer market outcomes, using their large purchasing power, in directions favorable to achieving overarching (environmental) policy objectives. These objectives can relate to environmental and social impacts, which are embraced under the United Nation's Sustainable Development Goals (SDGs) [37]. Essentially, GPP creates an effective top-down value chain measure by imposing requirements on contracting firms and their respective suppliers to pressure them into reducing their environmental and social impact [13][67]. To do this, governments mainly leverage the three-stage administrative process, as delineated in Section 3.3 [67].

GPP has transformed into a multiple-objective instrument aiming to fulfill a primary need in such a way positive externalities are made possible [44]. By applying GPP, the influence of governments on the market can become quite substantial [69]. Research finds that GPP can influence production and consumption, creating demand for green alternatives and expanding their market [14][27][32][41][42][63]. Additionally, it promotes innovation and advocates the adoption of environmental standards through private-sector investments [13][60].

Kumar (2022) [42] notes that while the utilization of Public Procurement to lower environmental impacts was novel, the idea of using PP for strategic purposes was over a century old. Governments used it from the 19th century onward, predominantly implementing regulations emphasizing cost and competition to address unemployment, labor conditions, and labor opportunities for the disadvantaged [42].

As an environmental policy, GPP can either facilitate substitution, transformation, or a combination of both [45]. For it to be a substitution policy, public purchasers must substitute conventional 'brown' products/services with more eco-friendly alternatives [45]. To act as a transformation policy, it should encourage suppliers to change and invest in less harmful technologies to accommodate the green market [45].

Lundberg and Marklund (2018) [44] evaluated GPP against the guiding principles for effective environmental policies. The principles state that policies need to target the source of the problem directly, warrant a one-to-one alignment between each instrument and its objective (Tinbergen Rule), and ensure independence between the objectives and instruments [44]. Their research finds that GPP, within the EU, struggles to satisfy these principles, especially because simultaneously addressing multiple objectives is a frequent practice, but it risks undermining the successful achievement of the core objective, namely to obtain the best value for public money through transparent, efficient, and non-discriminatory competition [44].

In contrast, a well-aligned policy example is a carbon tax. This instrument directly targets the environmental externality (carbon dioxide emissions), aligns one tool with one objective (emission reduction), and avoids dilution through additional goals, thereby fully adhering to the principles outlined by Lundberg and Marklund (2018) [44].

As mentioned before, the EU believes firmly in the capacities and potential of GPP. Existing literature acknowledges the potential power it can exert if properly implemented and utilized as a policy tool. However, several studies have raised critical questions and pointed out (potential) limitations of GPP.

Lundberg et al. (2015) [45] state that if GPP fails in its objective and environmental concerns within public contracts provide limited environmental gain but increase prices, two major issues come about. First, price-sensitive private consumers would opt for 'brown' alternatives. Second, the increased pricing under the guise of environmental improvement could introduce greenwashing, where suppliers falsely claim to be sustainable or exaggerate their sustainability efforts to mislead public authorities and charge higher prices. Another issue concerns the alignment of the desired environmental objective and the specified green criteria within a procurement. In their study of 337 procurement auctions, Lundberg et al. (2015) [45] were not able to identify what specific objective the listed green criteria sought to address once. The general conclusion from Lundberg et al. (2015) [45] and Lundberg and Marklund (2018) [44] is that the design of GPP has limited impact or no potential to function as an environmental policy instrument because of its limited perceived impact on supplier participation.

Lundberg and Marklund (2018) [44] find that GPP can better be described as an administrative tool to address environmental issues that sets specific rules or requirements, contrasting economic tools like taxes and subsidies. Generally, it is not the most cost-effective option, as it targets environmental goals indirectly through procurement rules rather than addressing pollution at its source. Nevertheless, it can be viable when preventing serious harm or long-term damage might outweigh the extra costs [44].

For the construction sector, (G)PP as a single instrument to address multiple objectives (building functionality, social impact, and ecological impact) is deemed to be an inconsistent policy system because the primary goal of completing the construction project could take precedence, leading to other aspects being deprioritized [44]. Lundberg and Marklund (2018) [44] underscore the need for complementary instruments to address broader goals within public construction projects.

Globally, widespread adoption of green procurement practices where environmental factors are prioritized in awarding public contracts has been observed [14]. Most of the EU Member States have already committed to applying and promoting GPP [13]. However, early on, uptake has been lower than expected [48][55]. This is partly due to the reasons presented in Section 3.1 but also because supplier participation is optional, which distinguishes GPP from more traditional environmental policy instruments, such as emissions taxes [45]. Suppliers' decisions to enter into GPP contracts are dependent on their expected pay-off; thus, if so-called 'brown' contracts are more favorable, suppliers can opt not to partake in green practices [45]. Another possible reason for the lower uptake is that even though the procurement directives allow public purchasers to use environmental criteria as part of their decision, it is not guaranteed that these authorities will successfully implement these criteria or even reward contracts to the greenest tenders [52].

While GPP is often presented as a promising tool for achieving environmental objectives, the findings discussed above suggest that its practical application faces significant challenges. Issues related to cost-effectiveness, unclear alignment between objectives and criteria, limited supplier participation, and inconsistent implementation across member states have raised concerns about its actual impact. These limitations highlight the need for a more realistic assessment of GPP's capabilities and for the development of complementary instruments to achieve environmental policy goals effectively.

3.5 Challenges in Translating Ambitions into Implementation

GPP is seen as an innovative policy tool, sparking ambitious initiatives and interest among policymakers and researchers alike [34]. Including environmental criteria in PP contracts can reduce negative environmental impacts [49]. However, the potential remains partially realized with its implementation coming across several obstacles [51]. The general lack of well-defined criteria has been heavily criticized [32][48]. Sapir et al. (2022) [59] add that the complexity of linking the criteria to the correct subjects[3] and unambiguously describing the environmental criteria are other barriers to proper implementation. On top of this, there is the fear that including environmental criteria has the

[3] Example given in Section 3.3

unwanted side effect of favoring certain companies and excluding others from the bidding process [67].

To function as effectively as possible, environmental policies, such as GPP, should be designed to directly target the source of the problem. Unlike more coercive instruments like carbon taxes, GPP's effect is often diluted due to design, adoption, and operational challenges [44]. However, this lack of coerciveness may also explain its political acceptability, as it allows policymakers to promote sustainability goals without imposing mandatory regulations on market actors. Additionally, Hafsa et al. (2021) [37] estimated that the true potential of PP to drive sustainability is underestimated due to incomplete data, possibly restricting the attractiveness of governments to uptake sustainable procurement. This section will address identified barriers to better understand why the EU's ambitions have poorly translated into practice.

Despite its perceived advantages and a growing list of tools, criteria, and success stories, the uptake of GPP among EU Member States has been slow and inconsistent [41][50][54]. Note that only four countries achieved the EU's 2010 target of 50% of all procurement procedures, including core GPP criteria, followed by eleven countries that achieved figures between 20% and 40% and the remaining twelve countries where the uptake was less than 20% [41]. In Nilsson Lewis et al. (2023) [50], an exposition into the PP landscape of eight EU Member States revealed that GPP was increasingly perceived as a key instrument, but implementation still remained fragmented. In 2020, only 1% of procurements in Poland and 4% in Estonia were considered green, while Sweden and the Netherlands achieved significantly higher rates at 58% and 67%, respectively [50].

The voluntary nature of the 2014 directives (see Section 3.1) and GPP's contingency on supplier participation (see Section 3.4) are hampering more prominent usage [45][54][57].

Some studies suggest that voluntary participation may reduce competition; suppliers may opt to avoid GPP auctions if green criteria require significant investments, leading to fewer bidders and higher prices [14][45]. Lundberg et al. (2015) [45], using detailed Swedish procurement data, find only weak statistical evidence for this effect, but highlight that the structure of GPP, especially the use of mandatory green criteria, can create entry barriers that potentially reduce bidder participation. Building on this, Drake et al. (2024) [14] similarly finds that competitive intensity tends to decline over time in green tenders, likely due to increased administrative burdens and complexity. Reflecting the general challenge of balancing short-term competitiveness with long-term environmental policy goals. Suppliers are likely to pass forward their high adjustment costs to the public sector, with stringent criteria often resulting in higher bid prices [44]. Additionally, too little binding requirements exist to motivate governments to practice it [50]. Also, the lack of appropriate regulations at both the national and international levels enables the scattered approach with limited connections between (inter)national and local procurement practices [15][50].

At the organizational and operational level, several barriers hamper the implementation. These can be grouped into four categories: knowledge and skills gaps, motivation, organizational opportunity, and resistance to change [3][36][41][50].

With a limited understanding of GPP practices, procurement officers can impede adoption by choosing not to apply environmental criteria at all, despite the legal possibility to do so[41]. The lack of knowledge and the increased complexity of sustainable procurement activities drive officers to make safe and traditional choices [36]. One of the most critical issues is identifying suitable environmental criteria and assigning appropriate weights to reflect their importance. However, there is currently a lack of standardization in how these criteria are selected, scored, and weighted. In practice, the bid price was frequently observed to dominate as the primary concern, causing environmental criteria to be deprioritized, undermining sustainability implementation [52]. Methodologically, challenges concerning poorly defined environmental criteria and limited knowledge and use of tools such as cost-benefit analysis (CBA) and life cycle costing (LCC) require standardization and better integration to improve GPP implementation [3]. The need for more awareness and clear guidance can be fulfilled by harmonizing processes, data, and ways of tracking impacts [50]. A partial step in this direction is the recently adopted Ecodesign for Sustainable Products Regulation (ESPR), which promotes harmonization by setting common sustainability requirements and enabling better data tracking through measures such as the digital product passport [23].

Motivation is a critical aspect of successful implementation, either positively or negatively. Grandia and Voncken (2019) [36] found that public procurers who are not intrinsically motivated to pursue sustainability tend to implement fewer sustainable practices compared to those with a stalwart personal commitment. Without intrinsic motivation, efforts to adopt and implement GPP are less likely to succeed. This is further compounded by the lack of external incentives to encourage procurement officers, resulting in apathy to pursue GPP practices [50].

Even if a sense of motivation is present and procurers are knowledgeable, the lack of organizational support could hinder execution. In order to achieve successful implementation, an environment that provides flexibility, fosters innovation, ensures sufficient resources, and has supportive leadership is needed [36]. However, lack of funding and capable staff, in combination with time pressure, remain delimiters within organizations [41][50].

Public sector inertia and reluctance to change traditional and proven procurement processes limit the willingness to change [3][41]. Adding to this are misconceptions about GPP, partly due to the lack of knowledge. In particular, the perception of higher costs has been observed to increase resistance [41]. This belief causes procurement officers to prioritize short-term budgets over long-term benefits [41]. Another driver behind this choice could be that procurement officers often work within budgeting systems that focus on short-term savings, making it difficult to justify higher upfront costs. Nillson Lewis et al. (2023) [50] propose fostering collaboration and networking among market actors and procurement officers to overcome this reluctance.

To successfully translate ambitions into reality, green criteria must be made mandatory, align with specific environmental objectives, and incentivize traditional 'brown' suppliers to adopt greener standards while ensuring the public sector is willing and capable of absorbing associated price premiums [14]. However, in practice, specifically in sectors like construction, where uncertainties exist surrounding costs and pay-offs, bidding decisions quickly increase in complexity [4]. Additionally, best practices are often negated when nearing the fiscal year-end, with remaining budgets 'needed' to be spent. Resulting in rushed, lower-quality procurement, further complicating market dynamics [57].

The **Institutional Theory** elaborated by DiMaggio and Powell in 1983 [56] provides a relevant framework for explaining the (challenged) adoption of GPP in the EU. The Institutional Theory states that three institutional pressures are shaping the adoption of sustainable procurement: coercive, mimetic, and normative [56].

- **Coercive Isomorphism** arises when influential stakeholders (the European Commission) impose regulations and policies [56]. However, the voluntary nature of the GPP directives weakens this pressure, leading to inconsistent adoption and misalignment with environmental objectives [50].
- **Mimetic Isomorphism** is when organizations imitate the path to success of others [56]. This pressure can affect both the supply and demand side within public procurement, but the lack of standardization and transparency regarding outcomes currently limits the influence [3].
- **Normative Isomorphism** relates to pressures from external stakeholders stemming from their demand for what they constitute as proper and legitimate behavior [56]. While public expectations increasingly favor sustainable procurement, this pressure may be less influential in practice

if buyers lack the support, training, or institutional frameworks to translate those norms into action[36][41].

3.6 How To Identify ”Green” Contracts?

The Public Procurement Directive 2014/24/EU encourages the inclusion of environmental criteria, among other possible non-price criteria such as social or innovation-related considerations, at the different phases of the tendering process, as described in Section 3.3 [38][46]. Including GPP criteria allows public authorities to address environmental concerns along the life cycle of their procurements, as well as bestowing a groundwork for selecting tenderers according to their environmental competencies [11].

However, although the 2004 directives gave examples, they did not dictate how environmental criteria should be constructed [52]. This resulted in much freedom in formulating and interpreting criteria and in how the 'greenness' of tenders was measured [52].

Palmujoki et al. (2010) [51] expressed a call for clearer, more detailed, and standardized environmental criteria to avoid giving the contracting authorities unrestricted freedom of choice. The 2014 directives answered this call and further refined the concept of green contracts with more explicit examples and guidelines [14].

It was mandated that criteria directly related to the subject matter of the contract—focusing on either the composition of the product/service or address the production/delivery, not broad corporate policies—and adhered to the general principles set down in the EU Directives (Directive 2014/24/EU, Directive 2014/25/EU) [14][50]. One such principle obliges procurers to accept equivalent proof of satisfied requirements provided by suppliers to avoid the obstruction of competition [50]. Nevertheless, considerable flexibility in determining environmental criteria still exists [3][14].

To improve transparency and competitiveness in the tendering process, the European Supreme Tribunal of Justice (ESTJ) ordered that environmental criteria must be clearly specified and measurable, disregarding general and immeasurable ones, to ensure transparency and equal treatment of tenders [33][51].

In legal terms, ”measurable” refers to criteria that can be objectively verified using standardized or widely accepted methods (e.g., certifications, emissions metrics, recyclability percentages). Nevertheless, the interpretation of ”measurable” has been shaped by both statutory law, such as the Public Procurement Directive 2014/24/EU, and case law. Judicial decisions have emphasized that while criteria must enable transparent and objective assessment, they should not impose unreasonable burdens or costs on bidders. This balance ensures that the criteria remain competitive while meeting the required legal standards for clarity and enforceability.

For the empirical part of this research, identifying whether a contract is 'green' is essential. This section will review how existing literature approaches this aspect, laying the foundation for how this study will handle this facet.

The public body's preliminary decisions are solely responsible for a contract's 'greenness' [3]. Ahmed et al. (2024) [3] note that due to the lack of distinguishing factors between traditional and green procurement decisions, determining if a procurement is green involves analyzing the entire procurement system. However, several attributes and decisions are prominently observed in GPP, including preliminary market consultations to better understand market capabilities, known as market dialogue, green criteria, Eco-labels, and third-party certifications [3].

In Palmujoki et al. (2010) [51], procurements were classified as green when the final contract included at least one environmental criterion. While this is a relatively low threshold, it reflects the broader understanding that GPP does not require the use of MEAT as the award method [15]. Environmental criteria can be present across multiple procurement components without being directly part of the evaluation formula, as described by Testa et al. (2016) [63]. This explains why tenders evaluated on price alone may still be considered green if they contain environmental elements elsewhere in the process. Keaveny and Butler (2014) [41] apply a stricter standard by warranting that multiple EU core criteria (within the construction sector) be present. These include, among others, energy efficiency, carbon reduction, sustainable materials, waste management, and life-cycle assessment (LCA) [3][12]. Sustainable materials are defined as materials that are durable, reusable, or recyclable, incorporate recycled content, and are sourced locally [1].

Criteria can be divided into two levels of rigor: core and comprehensive [11][12][50]. Core criteria are designed to be easily implementable by both contracting authorities and suppliers across Member States and aim to make a (significant) positive environmental impact with minimal additional effort or cost [11][12][50]. Comprehensive criteria are for those public bodies pursuing a higher environmental ambition that could require additional verification and administrative effort, and possibly higher costs [11][12][50]. In other research, the distinction was made between regular and well-defined environmental criteria, the latter referring to a criterion for which the procurer has provided a detailed description of how it must be fulfilled and verified [32].

GPP involves the introduction of environmental standards for technical aspects, award criteria, and performance clauses [15]. The green criteria act as mandatory qualification criteria to evaluate bids [45]. The criteria can be associated with the intrinsic qualities of the contract or relate to the production/delivery of the product/service [45]. Within sustainable construction, bids can be commonly evaluated on emission reduction, energy performance, environmental certificates, waste recovery, environmental qualifications of workers, efficient resource utilization, and life-cycle costs [15][50].

However, MEAT is not a prerequisite for GPP [15], both because environmental considerations can be integrated into other parts of the procurement process [63], and because MEAT may also be used to prioritize non-environmental criteria making it an insufficient indicator of "greenness" on its own. Additionally, Džupka and Kubák (2020) [15] argue that green criteria are often not used as qualifiers but rather incorporated into conventional criteria, because contracting authorities tend to avoid prioritizing green criteria over traditional ones. This phenomenon could potentially explain why the amount of green criteria in the call for tenders is double as high as in the final contracts, as mentioned in Palmujoki et al. (2010) [51].

Research into the definition and application of environmental criteria, especially in the construction sector, leaves much to be desired [33]. In Fuentes-Bargues et al. (2017) [33], four large criteria groups are proposed for awarding contracts in the construction sector. The first relates to the energy efficiency of both buildings as a whole and installed equipment. The second group includes sustainable construction materials, with LCA as a suggested tool for proper selection. The final two groups are measures designed to facilitate responsible water usage (collect rainwater, groundwater infiltration) and waste management (reduction, reuse, and recycle).

The analysis of Swedish cleaning service procurements identified 28 different environmental criteria, which were reduced to six categorized variables to manage dimensionality and streamline analysis [45]. The first category encompassed criteria related to environmental management frameworks, certifications, and standards such as ISO 14000. These certifications reflect a supplier's general environmental commitment, raising concerns about the appropriateness of such identified criteria, particularly when the possibility exists that they were not directly linked to the subject matter of the contract. The second category focused on ecological labeling requirements for cleaning products, such as EU Eco-label certification. Another category addressed vehicle emissions, including criteria such as fuel efficiency, electric/hybrid vehicles, and eco-driving practices. The chemicals category encompasses all criteria relating to compliance with chemical regulations (REACH directive, Swedish

Environmental Code). A separate category was created for monitoring activities. The final category, dubbed 'Other Environmental Criteria', consisted of broad requirements, such as recommendations from the Swedish Environmental Management Council.

Existing literature highlights the wide variation in how environmental criteria can be formulated and applied, emphasizing the challenges of determining what qualifies as "green" in practice. These insights provide a foundation for formulating a practical approach to identifying green contracts within EU construction procurement. Section 4.2 will provide the tailored identification method used in this thesis.

3.7 Measuring Competition in Public Procurement: Approaches and Proxies

To answer the prerequisite for RQ2: "How can bidding competitiveness be quantified using available data?", existing literature was analyzed to provide inspiration and guidance. The European Union defines a market where mutually independent businesses engage in the same activity and contend with one another to attract consumers as a free competition market [28]. Competition thus means that businesses are subjected to competitive pressures from others. If competition is effective, it creates a level playing field for businesses while yielding consumer benefits such as lower prices, increased quality, innovation, and product diversity [28].

The Single Market Scoreboard—a tool developed to monitor and evaluate the functioning and performance of the EU's Single Market—provides insight into the competition levels of PP, albeit simplified, by several indicators [17]. Tátrai et al. (2024) [65] identified the *Single Bid Indicator* as being the most important one from the viewpoint of competition. A high proportion of contracts awarded with a single bidder signifies lower competition; logically, more bidders are better, as this means public buyers have more options and can get better value for money [17]. In 2022, the proportion of contracts awarded with a single bidder rose to the highest level in over a decade [17]. Primarily, Central and Eastern European countries perform unsatisfactorily, with Poland and Slovenia even exceeding the 50% mark, well above the 10% good practice threshold. Other indicators that could serve as proxies for competition relate to SMEs, such as contracts with SME participation and SME bids. With most EU companies falling under this classification, higher scores are better. A low score could indicate that potential market barriers that prevent SME participation are present [17].

Researchers commonly use the number of bids as a proxy to measure the intensity of competition [65]. Drake et al. (2024) [14] found a nonlinear and negative relationship between the number of bidders and the size of the winning bid, suggesting that lower costs can be incurred by increasing competition. However, while the number of bidders and the level of competition are closely linked, this relationship is not exclusive, as wide-ranging market dynamics could play significant roles.

Competition is inherently endogenous, shaped by the elements within the procurement system (regulations, call characteristics, and the design of procedures) [14]. Nevertheless, Tátrai et al. (2024) [65] identified literature gaps regarding the tender characteristics and decision-making procedures that influence competition intensity.

Some identified competition-strengthening factors are contract value, contract duration, the nature of the tender procedure, and the awarding criteria [65]. A balanced estimated contract value attracts the most bidders, while extremely small or large ones deter participation. Duration-wise, more extended contracts are most appealing. Regarding the tender procedure, Tátrai et al. (2024) [65] find that elements such as dividing contracts into lots and utilizing negotiated procedures are associated with a higher likelihood of receiving multiple bids, and fewer single-bid submissions. While open procedures are generally seen as the default for promoting wide competition, the findings suggest that negotiation-based approaches may, in certain contexts, perform equally well or even better in attracting bidder participation. Lastly, using the lowest-price criterion to award contracts is prone to attracting more bids than the MEAT approach, which runs counter to the EU's sustainability goals [65].

Existing literature provides both conceptual clarity and empirical guidance on how competition in public procurement can be measured. Providing aid for selecting an appropriate proxy variable within this thesis. Section 4.4 will outline how bidding competitiveness was operationalized based on feasibility and relevance to the dataset at hand.

3.8 Green Public Procurement in the Construction Sector

Construction is estimated to represent around 5,5% [3] to over 10% [32][41] of the EU's GDP and employ roughly 7% of the available workforce, making it a sector of high importance [32]. The public sector can be an important lever within the construction sector to drive decarbonization by providing market demand, with public purchases accounting for 40-60% of global concrete consumption and 20-30% of the construction industry's revenues [50].

PP in this sector consists of three segments: works—constituting the construction of buildings, large infrastructures, or reforms that result from the construction process; services—operations performed by technicians who are involved in the process, such as architects, engineers, cleaning companies or security companies; and Products—encompassing elements or materials used to build the works or to facilitate the service. [33].

Construction and its related activities exert an immense strain on the environment [33][32]. The environmental harms (pollution and global warming) caused by the industry are detrimental and very apparent [4]. The industry is characterized by intense, excessive usage of natural resources [32][54]. It is estimated that roughly a third of the available amounts are extracted to be used in construction [32], and within the EU, this figure lies around 50% [18]. Additional characteristics are waste generation and enormous energy demands [32][54].

Studies show that the construction industry is responsible for over a third of the global GHG emissions [4]. In the EU, it accounts for an estimated 5–12% of total emissions [18]. The industry constitutes about 12% of GHG emissions brought forth by global public procurements [50]. In 2018, 39% of global energy and process-related carbon dioxide emissions were attributed to construction activities [50]. In 2022, this was estimated to be 37% [3]. A 2013 study of the EU construction industry found that the industry was responsible for 40% of European energy use and 36% of EU carbon emissions [41]. At a global level, around 54% of solid waste is produced as a result of construction activities [4]. Within the EU, these numbers are lower, between 35% and 40% of all produced waste [8][18]. Binshakir et al. (2023) [4] synthesize other adverse environmental effects caused by construction activities. Such as potential impacts on wildlife and natural features, soil and ground contamination, water contamination (surface and underground) and depletion. Also, the production of dust, noise, vibrations, and odors.

All these negative externalities make sustainable management of the construction industry and its produced waste an important target for the EU [8]. Key legislation such as the Energy Performance of Buildings Directive (EPBD) [24] and the Construction Products Regulation (CPR) [20] aim to raise environmental standards across the sector. The EPBD sets minimum energy performance requirements for buildings and promotes nearly zero-energy buildings across the EU [24]. The CPR, on the other hand, harmonizes rules for construction products and requires manufacturers to disclose the environmental impact of their products, often through Environmental Product Declarations (EPDs[4])

[4] EPDs are standardized documents that present quantified environmental data (carbon emissions, resource use, end-of-life impact) based on Life Cycle Assessment (LCA).

[20]. Therefore, adopting GPP in the construction sector is essential to reducing unwanted impacts [33][38].

Within this context, GPP is a complementary tool, aiming to reinforce the regulatory objectives through demand-side pressures. Nillson Lewis et al. (2023) [50] explain that in order to achieve the goal of reducing GHG emissions by 55% (compared to 1990 levels) before 2030, focusing on industries like construction, where decarbonization requires significant investments and disruptive shifts, should be the EU's priority to make the greatest environmental impact and support quicker decarbonization of heavy emitting industries.

Nevertheless, research and regulation on GPP in the sector have been insufficient [32]. Interest in GPP practices has increased, with perceived worldwide practice [14] and the incorporation in National Action Plans (NAP)—strategic policy documents developed by governments to outline specific goals, priorities, and actions within a particular area—across the EU [3]. However, in 2019, utilization of environmental criteria within construction procurements remained below the EU-wide average across all procurement sectors [32]. In a recently published work, Han et al. (2024) [38] claim that adequate integration of GPP is still lagging. Ahmed et al. (2024) [3] observed that the implementation of GPP in construction work and services is not as extensive as its proportion in overall public procurement. A reason for this is the inertia of the construction industry, with the EU 27 report highlighting an exceptionally low uptake of GPP, with only 3% of construction contracts including all applicable core green criteria, as defined in the relevant guidance for the sector [3].

Nilsson Lewis (2023) [50] provides a detailed elaboration on voluntary criteria for two subdivisions of the construction sector. Road design, construction and maintenance on the one hand and office building construction on the other. For office building construction, the focus lies on providing space for various service providers while limiting negative environmental impacts. The environmental impacts of roads include vehicle fuel consumption and emissions, as well as the risk of biodiversity loss. The scope of EU GPP criteria for this subdivision of construction considers three types of procurements: road construction—preparation and building of a road; road maintenance—actions related to maintaining and restoring the service level of roads; road reconstruction—upgrading road sections. The procurement of roads encompasses six stages—preliminary scoping, detailed design and performance requirements, construction, use of the road, maintenance, and decommissioning—in which green criteria should be incorporated. For office building construction, the scope includes building new buildings, as well as renovating existing ones, but parking lots are considered to be outside of the scope. Here, a total of eight stages—preliminary scoping, detailed design and applications for permits, site preparation works (strip-out and demolition), construction of the building or major renovation works, installation and supply of energy systems and services, completion and handover, facilities management, and post-occupancy assessment—exist where GPP can occur. For both subdivisions, examples of EU GPP criteria are given in Table 1. This table highlights the relationship and difference between general environmental requirements and specific award criteria.

Achieving environmentally sustainable construction is about striking a balance between the environmental aspects of construction and the costs to optimize the benefits [41]. To achieve this, environmental criteria must be incorporated into construction contracts [41]. Simcoe and Toffel (2014) [60] stipulate that public sector GPP policies could also promote the diffusion of greener practices within the private sector. This can be done via three mechanisms. Firstly, governments could raise awareness about green construction benefits, potentially increasing private sector demand [60]. Second, public sector adoption of GPP might lower prices for green building inputs through a combination of scale economies, reduced entry barriers for suppliers, and learning effects [60]. Thirdly, the public sector could act as a temporary intermediary player to solve a possible coordination problem within the sector [60]. For example, private-sector developers may wait for suppliers to invest in green building technologies and skills, while at the same time, suppliers are waiting for a substantial demand for these green practices. Governments can thus create a guaranteed demand for suppliers to prompt the growth of this specialized market.

Challenges, however, do exist. The green premium (higher upfront costs for low-carbon technologies and materials) associated with green contracts might scare suppliers and procurers [50]. Additionally, sustainable projects have more restraints and complications than traditional construction projects [4]. Combining this with the lack of green construction experience significantly raises the complexity of the bidding decision [4]. This decision is of crucial importance for contractors, as selecting a potential project to bid for influences the construction companies' future success [4].

Binshakir et al. (2023) [4] analyzed forty factors that could influence a contractor's decision to enter a green procurement auction. Their analysis revealed that financial aspects—payment history of procuring agency, financial capability of supplier and procurer, and required investments; project characteristics—project risks, size, type, and complexity; and experience were among the most influential factors [4]. Again, the higher financial burden associated with sustainable construction acts as a barrier to supplier adoption of GPP.

Lundberg and Marklund (2018) [44] question GPP's ability to constitute a consistent policy system in the construction industry. The primary purpose of PP in construction is to fulfill a contract authority's need for a specific type of infrastructure. Therefore, the main binding objective is to construct said infrastructure, independent of environmental objectives that may accompany it; the reverse, however, is not true [44]. This dependency can be concretized as follows: a public body could commission the construction of a building without environmental criteria, but it cannot require the inclusion of criteria before procurement of the construction is undertaken. Hence, GPP in construction violates the third guiding principle of consistent policy (mutual independence) [45]. The second guiding principle (Tinbergen Rule), which states that the number of policy instruments must match the number of policy objectives to achieve all objectives effectively, is also not fulfilled [44]. Imagine if a public agency wants to procure a new office building. This constitutes their primary objective (Objective X: construct a functional office building). If the agency also wants to reduce the building's energy usage by incorporating LED lights (Objective Y) and only wants to use local, sustainable materials (Objective Z), issues arise. If two objectives (X and Y) need to be achieved, the procurement auction can only offer one unique solution. If a third objective is added (Z), achieving all objectives becomes impossible. At least one objective must be compromised. For instance, requiring all building materials to be local might limit the energy efficiency of the LED lights.

The effectiveness of GPP in the construction sector is obstructed by industry inertia and financial concerns among suppliers, but given the environmental impact of the sector, change is necessary. However, concerns about GPP being the correct instrument to solve these issues still exist.

4 Methodology and Data

This section outlines the methodological approach used to evaluate how including environmental award criteria affects procurement outcomes in construction sector tenders. As mentioned in Section 2, this study is guided by three research questions: the extent of environmental criteria usage, their effect on bidding competitiveness, and their effect on project cost. These questions shaped the empirical strategy presented below, which harnesses a Difference-in-Differences framework to assess causality.

Section 4.1 introduces the overall research design, and several methodological options are considered in light of the study's causal research objectives. Section 4.2 provides a detailed description of the dataset and the key data-related challenges that shaped the methodological approach. The

Environmental Requirement and Award Criteria		
Environmental Requirement	**Award Criteria**	**Example**
Road Construction Voluntary EU GPP Criteria		
Tenderer experience	Ability of the tenderer	Previous experience in similar projects
Technical performance	Pavement-vehicle interaction	Earthworks and groundworks, rolling resistance
	Resource-efficient construction	Materials choice, production and transportation
	Maintenance and rehabilitation strategies	Preventive maintenance plans
Environmental impact	Congestion	Traffic mitigation plan
	Water and habitat conservation	Drainage
	Noise	Noise in construction process and road noise
Office Building Construction Voluntary EU GPP Criteria		
Emissions	Low- or zero-carbon energy sources	Installing solar panels
	Building life cycle Global Warming Potential (GWP)	Using software tools to calculate the total CO_2 emissions over the building's life cycle
	Performance requirements for CO_2 emissions from the transportation of aggregates	Sourcing construction materials from nearby quarries
Energy	Minimum energy performance requirements	Installing energy-efficient lighting (LED)
Material sourcing	Incorporation of recycled or re-used content in concrete and masonry	Reusing bricks and tiles salvaged from demolition
Environmental performance assessment	Performance of the main building elements: Aggregation of environmental product declarations	Selecting building materials with verified EPDs
	Performance of the main building elements: Carrying out of a life cycle assessment	Conducting an LCA to compare the environmental impact of using wood vs. concrete for the structural framework

Table 1: Voluntary EU GPP criteria for Road and Office Building Construction, adapted from Nillson Lewis et al. (2023) [50]

rationale for selecting the final method is presented in Section 4.3, which also includes a critical reflection on its strengths, limitations, and relevant assumptions. Finally, Section 4.4 describes the operational implementation of the selected method, including treatment definitions, model setup, and assumption testing.

4.1 Quasi-Experimental Design

The nature of the research questions posed in this thesis, in particular RQ2 and RQ3, necessitates a causal analysis rather than a purely descriptive or correlational approach. These questions intend to approximate the impact of environmental criteria on the desired project outcomes, not just to detect associations.

Ideally, answering such questions would involve a Randomized Controlled Trial (RCT) in which public authorities randomly assign award criteria to procurement contracts, thereby isolating causal effects. However, in real-world procurement settings, especially at the European level, random assignment is neither ethical nor feasible.

Public contracts must treat all bidders equally. Randomly assigning award criteria could reduce fairness. Another ethical concern could be the potential suboptimal allocation of taxpayer money. Regarding feasibility, researchers have no control over how public tenders are designed, as these procedures are governed by legal frameworks. The random assignment of award criteria across contracts is, therefore, not possible. Additionally, experimental implementation is highly impractical due to the decentralized nature of procurement across countries and authorities.

Public contracting authorities choose award criteria based on project, political, or institutional factors, which may also affect the outcomes of interest (e.g., more complex or high-profile projects might be more likely to use environmental criteria and cost more). This introduces selection bias and confounding issues that a simple comparison of averages between treated and untreated contracts would not address. Therefore, a quasi-experimental approach is necessary to approximate a counterfactual scenario, entailing what would have happened to the same contract had it not included environmental criteria?

RQ2 (effect on bidding competitiveness) and RQ3 (effect on project cost) aim to estimate causal effects, not just statistical associations. A descriptive approach cannot rule out that outcome differences are due to confounding variables (e.g., country, project size, contract complexity). Although RQ1 is primarily descriptive, it plays a critical role in supporting RQ2 and RQ3, as it helps assess whether using environmental criteria is sufficiently widespread and varied to allow for quasi-experimental analysis.

To address the causal character of RQ2 and RQ3, several quasi-experimental methodological strategies were explored and considered. However, the available data, scope, assumptions, and the policy itself presented several constraints limiting the utility, robustness, and feasibility of these methods.

Propensity Score Matching - Difference in Differences (PSM-DiD)

A common approach to measuring the effectiveness or impact of a specific intervention or regulation is the difference-in-differences method or DiD for short. DiD relies on data collected through repeated observations of the same units over time, meaning the data must be longitudinal (also known as panel data) [7]. In its simplest form, DiD requires data from two groups observed across two time periods. These are the pre-and post-treatment periods and the control and treatment groups. The control group consists of units that do not receive the treatment, while the treatment group consists of those who receive the treatment [5].

DiD is a counterfactual impact evaluation method, which means it compares what actually happened with what would have occurred in the absence of the treatment (in this case, the policy change in public procurement introduced in 2014) [7][16]. DiD estimates the causal effect by comparing outcome changes between treated and untreated units over time. It does so under the key assumption that the unobserved differences between the treatment and control groups would have remained constant in the absence of treatment [5][7]. This makes DiD especially useful in contexts where random assignment at the unit level is unfeasible.

A commonly used extension of this method is the Propensity Score Matching combined with Difference-in-Differences (PSM-DiD). This technique improves the comparability between treatment and control groups by matching units based on observable characteristics [16]. It does so by calculating a propensity score, which represents the conditional probability that a unit receives the treatment, given a set of observed covariates [58].

The propensity score can be interpreted as the likelihood that a unit would have received the treatment based on its observed characteristics. The more strongly the treatment is confounded with these covariates (i.e., the more treatment is related to other variables), the greater the bias in the estimated treatment effect if left uncorrected. PSM-DiD addresses this by attempting to reduce selection bias stemming from observable characteristics through matching (via PSM) while controlling for unobservable, time-invariant characteristics through DiD [16].

In this study, selection bias may arise if treated tenders (i.e., tenders that include environmental criteria) are systematically different from untreated tenders. These differences could stem from observable characteristics such as the country of the contracting authority, the contracting authority type (CAE), or the type of work performed. For instance, tenders in regions with stronger sustainability mandates may be more likely to include green award criteria. Matching based on these covariates using the PSM technique can reduce bias from such observable factors [16].

However, systemic differences may also arise from unobserved factors, such as the procurer's experience, the contracting authority's internal procurement culture, or the procurer's affinity with sustainability, none of which are directly captured in the dataset. Thus, even tenders that appear similar in observable terms could differ in how seriously environmental criteria are applied or enforced. This is where DiD complements PSM: by comparing changes in outcomes between matched treated and untreated tenders over time DiD helps to control for time-invariant unobserved heterogeneity [16].

Methodological Requirements and Assumptions

Theoretically, PSM-DiD aligns well with the structure of the available data. First, a clear temporal dimension is present, allowing for pre- and post-treatment comparisons. Second, it was possible to identify treatment and control groups based on the interpretation of specific value fields, namely CRIT_CODE and CRIT_CRITERIA, which indicate the use of environmental or other award criteria. In addition, several covariates were available to control for observable characteristics, supporting the assumptions underlying the matching process.

However, the correct implementation of the PSM-DiD approach relies on several critical assumptions and data-related requirements. The most critical assumption of the DiD component is the parallel trends assumption, which states that, in the absence of treatment, the treatment and control groups would have followed the same outcome trend over time [5][7].

If the equal trends assumption holds, it implies that the null hypothesis: "There is no difference in the trend of the outcome variable between the treatment and control groups during the pre-treatment period", cannot be rejected (i.e., the interaction term is not statistically significant at a desired significance level) then it is reasonable to attribute any post-treatment difference in outcomes to the treatment itself. If, in the absence of treatment, the average outcomes for both groups would have similar trends over time, one can approximate the average treatment effect for the treated subpopulation (ATT) by comparing the average change in outcomes experienced by the treated group to the average change in outcomes experienced by the control group [5]. While the assumption of equal trends cannot be directly proven, its validity can be assessed through visual inspection of pre-treatment trends and statistical testing using pre-treatment data.

A second important condition for the DiD approach is the assumption of no anticipation effects, which states that in the periods prior to the adoption of the policy, the average differences in outcomes between the treatment and control groups must be zero. In other words, the treatment should not have an effect before it is actually implemented [5]. This assumption is implicitly included in the assumption of parallel trends.

For the matching component of the PSM-DiD approach to be valid, the Conditional Independence Assumption (CIA) must hold. This means the treatment assignment is assumed to be as good as random after controlling for observed covariates. This assumption is only credible if the dataset contains informative and relevant variables that sufficiently capture the factors influencing treatment assignment. Therefore, the PSM-DiD method requires abundant data on the main characteristics of both treated and untreated units before and after treatment [16].

In addition, a clear understanding of the factors that influence treatment assignment is necessary to define an effective matching strategy. Without this insight, it becomes challenging to select appropriate covariates for the matching process or to interpret the resulting treatment effect as causal.

Lastly, while a strictly balanced panel is not required, DiD estimation assumes that each unit is observed in at least one pre-treatment and one post-treatment period. Without this, it becomes difficult to estimate changes in outcomes over time within each group in a robust manner.

Methodological and Data Limitations

Panel data is recommended to estimate the causal effect of, for example, environmental (green) criteria on procurement outcomes using PSM-DiD. However, panel tracking over time was not feasible in this study, as each call for tenders notice[5] was treated as a unique and independent observation, regardless of whether it originated from the same CAE. As such, tracking repeated behaviour over time could not be achieved naturally. Additionally, to address structural issues in the TED data, such as duplicate lots, missing or unstandardized lot identifiers, and unreliable links between lots and contract awards, the dataset was restructured at the notice level. By aggregating and deduplicating records at this level, data inconsistencies were reduced and the risk of misclassification or data loss was minimized. Multiple contract awards were therefore consolidated into their corresponding notice. A visual representation of the aggregation is presented in Appendix 7.13.

To still enable a comparison of units across time periods, the analysis was designed to match treated units with untreated units. Nonetheless, attempts to construct clearly defined control and treatment groups based on untreated countries, contracting authority types or unaffected CPV categories were unsuccessful. Therefore, an unconventional PSM approach was adopted to create a synthetic pre-treatment group.

Pre-policy "green" notices[6] were matched with similar "non-green" notices. These matched "non-green" notices were treated as the pre-policy treatment group, while unmatched non-green notices served as the pre-policy control group. The post-policy treatment group consisted of all "green" notices awarded after the policy change, and the post-policy control group included all "non-green" notices, or only those evaluated using the lowest price criterion[7].

[5] Note: While this approach refers to the call for tender level as the unit of analysis, the underlying data is extracted from Contract Award Notices (CANs). These CANs represent awarded responses to previously published calls for tenders (Contract Notices, CNs). In this context, the term calls for tenders level refers to the aggregation of contract awards reported in a single CAN, which in practice corresponds to a completed procurement procedure originally announced via a CN.

[6] Analysis revealed that "green" notices could appear before the policy change introduced by the 2014 Public Procurement Directive.

[7] This distinction depends on the chosen treatment definition: are non-green MEAT notices included or not? In Section 4.3 the need for this decision is elaborated on.

However, the matching quality was suboptimal. Diagnostics such as standardized mean differences, variance ratios, and empirical cumulative distribution functions (eCDFs) between treated and untreated groups were frequently outside of recommended thresholds, raising concerns about the comparability of the matched units and the robustness of any causal claims derived from this approach. Nearly all configurations tested also failed to satisfy the assumption of parallel trends. The full R output for all matching diagnostics is provided in Appendix 7.14.

Validating this assumption proved difficult due to the absence of panel tracking for individual units. Moreover, the data was heavily right-skewed, and the synthetic pre-treatment group consisted of substantially lower-valued contracts. Although log transformations and other adjustments were applied to reduce the tendency for treatment group contracts to be consistently matched with lower-value contracts, the issue remained unresolved.

Another drawback is that PSM requires large, diverse samples to ensure sufficient group overlap. In this study, such overlap was limited, leading to poor matches and reducing both the validity and generalizability of the findings [16].

In addition, data on key unobserved factors influencing the inclusion of environmental criteria, such as internal procurement culture, prior experience, sustainability preferences, or project duration, was unavailable, which undermines the credibility of the Conditional Independence Assumption (CIA). Since factors that influence both treatment assignment and outcomes but are not observed cannot be controlled for in the matching procedure, the risk of hidden bias remains [35]. While the DiD component is designed to differentiate out time-invariant unobserved heterogeneity, it is uncertain to what extent this effectively captures all relevant confounders in this context, particularly if unobserved factors vary over time or interact with the treatment effect.

Potential Solutions and Methodological Improvements

One potential solution to overcome the limitations of the current setup would be to search for a suitable control group outside of the EU, where similar policy changes have not occurred. However, given the scope and timeframe of this study, this approach was deemed unfeasible in consultation with the thesis supervisor.

Another alternative would be to shift the unit of analysis from the call for tender level to the Contracting Authority or Entity (CAE) level, which was ultimately the approach adopted. Enabling procurement behavior tracking at the authority level rather than the individual contract level.

However, this method is not feasible within the described PSM-DiD framework, as it requires a sufficient number of observations per CAE for each year to allow matching while accounting for year-specific effects. In practice, this condition is not consistently met, as contracting authorities are not obliged to publish tenders annually, and many appear only sporadically in the dataset. Additionally, the analysis would no longer capture outcomes linked to individual notices, but would instead focus on how procurement behavior changes over time at the level of each CAE. This shift makes it possible to compare green and non-green tenders issued by the same authority, helping to better identify how the adoption of environmental criteria affects procurement behavior. In this way, the research question shifts from assessing individual tender outcomes to examining the broader effects of Green Public Procurement and the use of environmental criteria over time.

While the CAE level was ultimately used in the final two-period DiD analysis, where all pre- and post-treatment years were consolidated, such consolidation was deemed unsuitable for the PSM-DiD design, which relied on matching within specific years to control for time-varying confounders.

Future research could address this limitation by expanding the observation window or integrating external datasets to capture additional procurement activity across years, provided that data is available and consistent across CAEs.

Additionally, the analysis would benefit from access to more detailed covariates, such as project duration, reasons for award or rejection, evaluation criteria weights, use of framework agreements, and whether the tender was part of a multi-lot setup. Another valuable indicator would be procurement expertise or capacity, which could be proxied by the number of tenders published per CAE per year.

While some of these variables were available in the dataset, they were often missing, incorrectly formatted, or ambiguously recorded, which made them unusable for reliable matching or analysis.

Staggered Difference-in-Differences

Another pitfall of the PSM-DiD approach had to do with a made design assumption. The nature of the policy intervention itself posed a fundamental issue. Although the PP directive was published in 2014, with a formal transposition deadline in April 2016, awarding contracts based on MEAT is not mandatory for CAEs from that point onwards. Adoption can occur gradually and voluntarily, meaning that the treatment is likely introduced at different points in time across entities, making a standard DiD setup with a fixed cut-off less appropriate.

The framework brought forth by Callaway and Sant'Anna (2021) [5] on staggered DiD was adopted and adapted to address this issue. This method allows treatment to be introduced at different times across units and accounts for treatment effect heterogeneity (various periods and regions). The framework enables estimating the so-called group-time average treatment effect on the treated *ATT(g,t)* for each group *g*, which denotes the treatment timing, at a time *t* [5].

The individual estimates are aggregated into a single $ATT_{\text{aggregate}}$ with the following formula [5] where the weight function $\omega(g,t)$ varies across groups and periods, predominantly determined by the group size.

$$ATT_{\text{aggregate}} = \sum_{g \in G} \sum_{t=2}^{\tau} \omega(g,t) \cdot ATT_{(g,t)}$$

This results in a causal estimate of the average effect of treatment on the treated, properly accounting for staggered treatment timing [5].

Methodological Requirements and Assumptions

As staggered DiD is a derivative of the standard DiD approach, both the parallel trends and no anticipation effects assumptions remain applicable and must be validated to ensure the credibility of the estimates. Other relevant assumptions and cleaning steps from the PSM-DiD approach were maintained and integrated into this updated design.

This research design attempted to use a methodology to track individual CAEs over time. Assigning each CAE with a unique identifier enabled the linkage of multiple call for tenders issued by the same entity across different years. Once more, multiple contract awards were consolidated into a single, unique call for tenders notice to address issues related to human input errors during data entry. As a result, every observation in the dataset now represents a single call for tender in a given year, linked with other calls placed by the same CAE, generating the desired longitudinal panel format.

This structure allowed insight into the broader tendering history of each CAE and enabled a more accurate and temporally grounded distinction between the treatment and control groups.

The control group consisted of CAEs that consistently used the lowest price criterion in their calls for tender throughout the 2012-2023 period and, as such, could be labeled as never adopting environmental or other non-price award criteria. To identify these CAEs, the following criteria were applied:

- The CAE must have only used the lowest price criterion across all observed call for tenders.
- The CAE must have a call for tenders observed in at least two years between 2012 and 2023.

- The CAE must have at least one call for tenders observed in both the pre- and post-treatment periods.

The treatment group includes all CAEs that transitioned from price-based evaluations in the pre-treatment period to incorporating green (or other non-price criteria, depending on the interpretation of the treatment) criteria at some point after introducing the 2014 directive. The following rules were applied:

- The first observed call for tenders for the CAE must have been awarded using the lowest price criterion.
- The first observed call for tenders must occur before 26/02/2014 to ensure a valid pre-treatment baseline.
- The first treated call for tenders must appear after or on 26/02/2014.
- The CAE must have a call for tenders observed in at least two years between 2012 and 2023.
- The CAE must have at least one call for tenders observed in both the pre- and post-treatment periods.

Applying these rules facilitated approaching the policy's heterogeneous nature more accurately, allowing insight into the policy uptake and potential treatment effects based on when CAEs began adopting green awarding criteria. Applying these changes warranted the use of staggered DiD to ensure that treatment is not conflated with pre-existing heterogeneity between entities.

An important assumption specific to the Callaway and Sant'Anna framework is that once a unit adopts the treatment, it remains treated from that point onward [5]. In this study, the tracked unit is the CAE. However, this assumption is not perfectly aligned with the data structure, as a CAE might award a contract using green criteria in 2016 but could revert to using the lowest price criterion for a call for tenders in a later year.

To reconcile this, an interpretive approach is used to define treatment. Once a CAE has awarded a contract using environmental criteria, it is assumed that a mindset shift has occurred within that organization or entity. From that point onward, the CAE is considered treated, as they have demonstrated a willingness to consider sustainability in procurement decisions. Even if subsequent calls for tenders do not explicitly include green criteria, it is reasonable to assume that sustainability has entered their decision-making process. However, this may not always be visible in the available metadata or choice of award criteria.

The following variables are needed to operationally implement the staggered DiD framework: A unique identifier corresponding to the level at which the unit is tracked. A time-invariant grouping variable indicating if a unit belongs to the control or treatment group. A time variable indicating the period in which each observation occurs. A variable indicating the period when the adoption of treatment occurred for each unit, this variable has a fixed value for each observation per unit. A time-varying grouping variable, this variable equals one if the period comes after or equals the period when the adoption of treatment occurred. In essence, it indicates which observations of a unit were before and after the unit implemented the treatment. A dependent outcome variable of interest. Optionally, covariates can be included to control for confounding factors.

Methodological Constraints and Possible Improvements

One of the main challenges encountered was the high degree of missing data in the panel structure: very few CAEs were observed every year between 2012 and 2023. Fortunately, the *did* R package by Callaway and Sant'Anna supports unbalanced panels, so the estimation procedure can still be carried out.

To reduce the gravity of missing observations, experimental configurations were run that aggregated observations into two-year periods, for example, grouping 2012–2013 as Period 1, 2014–2015 as Period 2, and so on. This lowered the proportion of missing periods for each CAE and made the event-study plots more interpretable.

However, Callaway and Sant'Anna (2021) [5] advise to only aggregate periods based on logic, not at random. The time aggregation used in this study follows a logical structure based on policy and external events. The years 2012–2013 are treated as the pre-policy baseline period, while 2014–2015 capture the immediate aftermath of the policy introduction. The period 2016–2017 includes the formal transposition deadline of the directive and the first full year of its implementation. 2018 and 2019 represent a stable post-implementation phase, without any major disruptions and prior to the COVID-19 pandemic. The 2020–2021 period reflects the pandemic years, during which procurement processes may have been significantly affected. The final two years are considered part of the recovery period, marking a gradual return to typical procurement activity.

In addition, another configuration attempted to impute missing data using K-Nearest Neighbors (KNN), both on its own and in combination with the two-year period grouping. The idea was to improve coverage for CAEs that only appear sparsely while preserving as much of the underlying structure as possible. However, one must remain cautious about how much imputation is acceptable in the context of causal inference, and these results were deemed as exploratory. In some configurations, the fraction of imputed observations was larger than the fraction of actual observed observations.

Whilst determining the best amount for k, the best-achieved values for the relative Root Mean Square Error for each target variable were 0.042 for the contract price, which is acceptable, and 0.5878 for the number of offers, which is likely not great.

As mentioned earlier, one core assumption of the Callaway and Sant'Anna framework is that once a unit is treated, it remains treated [5]. During data exploration, varying usage of award criteria across years, even within the same CAE, was observed. To resolve this, the first call for tenders awarded using green criteria was interpreted as a mindset shift.

This conceptual resolution enforces this treatment assumption. While it reasonably fits the policy context, discretion is advised, and future research must determine whether this interpretation is methodologically sound or considered too speculative.

With regard to validating the assumption of parallel trends, the interpretation of the results proved challenging. In several configurations, the test statistic was returned as zero, and while the confidence intervals for the pre-treatment periods typically included zero, they were often broad and imprecise. This presents a limitation in assessing the validity of the parallel trends assumption, as it is unclear whether such outcomes indicate genuine support for the assumption or merely reflect low statistical power.

A potential solution in future research would be to look for and increase the number of observations per group and time period, which may help narrow the confidence intervals and improve the precision of pre-trend estimates.

As previously mentioned in the context of the PSM-DiD method, this research design would also benefit from more detailed and reliably recorded covariates. Although many theoretically useful variables are present in the dataset, they were frequently missing, incorrectly formatted, or ambiguously recorded, rendering them unusable for robust outcome modeling.

Due to the combination of a high quantity of missing data, enigmatic treatment behavior, and difficulties in validating key assumptions such as parallel trends, the staggered DiD approach was ultimately deemed unsuitable for the final analysis. While the method offers strong theoretical advantages, its application in this study was limited by the quality and incomplete coverage of the available data.

4.2 Data Collection, Processing, and Limitations

Data Source and Scope

By analyzing European construction public procurement tender data, this research aims to answer the research questions formulated in Section 2. The raw dataset consists of twelve CAN datasets in CSV format, from 2012 to 2023, collected from the Tenders Electronic Daily (TED). It contains 9 560 681 observations and 75 value fields, before the initial cleaning and filtering. The data originates from unverified inputs from contracting authorities or entities across Europe, implying that incorrect or missing entries are not uncommon. Therefore, careful data management and cautious interpretation are essential.

This study exclusively uses Contract Award Notice (CAN) datasets, rather than Contract Notice (CN) datasets. While the latter includes all published calls for tenders, these may remain unanswered, meaning they lack observable outcomes. As a result, they are unsuitable for estimating effects on variables such as competition and contract value. Additionally, all relevant variables available in the CN dataset are present in the CAN dataset, making the inclusion of CNs redundant for this analysis.

The initial dataset is structured at the contract award notice level, where each observation reflects an outcome of a specific call for competition. A many-to-one relationship exists between contract awards (CAs) and their respective calls for tenders. Each CA is linked to a single call for tenders (identified by the ID_NOTICE_CAN variable), while a single call for tenders may result in multiple contract awards. The dataset also includes a substructure for lots, but this was not used in the analysis due to a lack of standardization in how lots are linked to CAs. This made it impossible to reliably interpret these value fields.

Cleaning and Preparing the Dataset

Focusing on the construction sector, a filter was applied based on the Common Procurement Vocabulary (CPV) code associated with each CA. All construction-related activities fall under the CPV code series "45XXXXXX-Y". Applying this filter yielded 748 344 observations, with 8 812 337 records excluded from the full dataset.

A rigorous and multi-stage cleaning process was undertaken to prepare the dataset for analysis. The first step involved filtering out observations that lacked a valid ID_AWARD or were marked as non-awarded, discontinued, or unsuccessful. Similarly, framework agreements and contracts with missing or zero values for the award price fields were excluded.

A central component of the analysis involved identifying "green" contracts. The CRIT_CRITERIA field, which contains unstructured text describing additional award criteria, was used as the basis for this identification. However, the field includes multilingual input, often with inconsistent formatting. To process this field, a Python script was used to translate the entries using the open-source LLM developed by the Language Technology Research Group at the University of Helsinki. Manual corrections were subsequently applied to fix incorrect or failed translations. A binary variable, is_green, was created and set to one if a predefined environmental keyword was found in the translated CRIT_CRITERIA and zero otherwise. The complete list can be found in Appendix 7.15.

Subsequent cleaning steps involved handling missing or inconsistent values across various variables. Many fields, such as CRIT_WEIGHTS, CRIT_PRICE_WEIGHT, and those related to lot-level granularity (e.g., ID_LOT, LOTS_NUMBER), were removed due to the lack of standardization, vague formatting, or excessive missingness. Contracts that included cancellation flags or unclear information on award outcomes were also excluded. Multiple fields were compared for price variables to create a single final_price_contract variable. The prioritization logic followed a hierarchy of available price fields, and contracts with missing or zero price values were dropped. The contract values were adjusted to 2023 euros using inflation factors to ensure comparability. The resulting adjusted prices were stored under the adjusted_poc variable.

To prepare for the analysis of project costs (RQ3), the data was further filtered to include only 'Works' contracts. Contracts categorized as 'Supplies' or 'Services' were removed. To align with EU thresholds that apply specifically to the construction sector, contracts below the €5 538 000 threshold were excluded (364 406 observations), and based on a visual inspection, an upper cap of €90 million was applied to remove extreme outliers. In addition, CAs that received more than 100 offers were considered atypical and were also removed.

To address issues related to human input errors during data entry, it is important to recognize that the TED data is provided *as-is*, with no standardized validation across CAEs in Europe. Because the data originates from thousands of decentralized public bodies who manually complete procurement notices, mistakes can easily occur, such as inconsistent formatting, typographical errors, or misinterpretation of fields. For example, one authority might enter the term "recycled materials" as an environmental criterion, while another uses a misspelled or incomplete version like "recyceld materials", making automated detection unreliable. Similarly, when multiple contract awards are issued under the same procurement notice, some authorities incorrectly entered the total contract value for each award, instead of dividing the value across individual contracts. These types of inconsistencies, often due to varying interpretations or lack of data entry guidance, required extensive preprocessing and rule-based harmonization to ensure the validity of the analysis.

To manage these issues, the dataset was restructured at the call for competition level (ID_NOTICE_CAN). Many call for tenders had multiple associated CA entries, and discrepancies could arise in fields like is_green, crit_code, or SME status. When such inconsistencies were detected, assumptions were made to harmonize the values. For instance, if any CA within a call for tender involved an SME, the entire call for tender was marked as involving an SME. Aggregated variables were created for each contract, such as the total number of offers, the number of offers from SMEs, the award date (split into year, month, day), and the consolidated green status. To ensure uniformity, only the first occurrence of each unique ID_NOTICE_CAN was retained, reducing the dataset to 60 753 observations.

Following these steps, the resulting dataset contained only observations with sufficient, verifiable information on contract value, award criteria, green content, and bidder characteristics. The only systematically missing variable was CRIT_CRITERIA, which is expected for contracts evaluated solely based on price (L contracts).

Standardization and Variable Transformation

In addition to the primary data cleaning process, several other necessary data transformation and enrichment steps were conducted to prepare the dataset for advanced analysis.

To identify individual Contracting Authorities or Entities (CAEs), a new variable CAE_ID was created. This was necessary due to the lack of standardization in CAE-related fields (e.g., name, address, town, postal code). A combination of string matching techniques and word searches was used to identify and unify entries that likely referred to the same CAE. Countries were grouped into five broader EU regions (North, West, South, Central, and East) to capture regional variation based on their geographical and political alignment. This classification was applied using the COUNTRY variable.

Simultaneously, CAE_TYPE values were consolidated into grouped categories: "National", "Regional", "International", "Sector-Specific", and "Unspecified". This transformation helped simplify analysis across procurement types. For both country and CAE type, the original variables were kept.

To incorporate broader contextual information and improve causal inference, five country-level macroeconomic indicators were added to the dataset as control variables. These covariates help isolate the effect of environmental award criteria by accounting for sectoral and economic trends unrelated to the treatment. Specifically, the SDG Index Score (sustainability performance) and the Spillover Score (externalization of negative impacts) control for a country's environmental and policy context. The annual growth of the Construction Producer Price Index (CPPI) adjusts for sector-specific inflation. Government Expenditure (as a percentage of GDP) captures national fiscal capacity. Lastly, to control for disruptions caused by the COVID-19 pandemic, monthly infection data (cases per capita, scaled by country) was merged based on year and month.

All observations for Liechtenstein (LI) were dropped, as they were only available in the later years of the dataset. Similarly observations for the United Kingdom (UK), Greece (GR), Switzerland (CH), and North Macedonia (MK) were removed from the dataset, as these were not contained in the COVID dataset when joining.

Treatment Assignment and Cohort Creation

Contracts labeled with CRIT_CODE = MEAT but without matching green criteria were excluded to ensure a consistent and focused classification of green versus non-green contracts. Since this study aims to estimate the effects of environmental criteria on procurement outcomes, including M-contracts that apply other non-price criteria, could introduce bias. These other MEAT factors, introduced by the 2014 directive, may independently affect outcomes such as price or competition. Therefore, retaining only M-contracts explicitly mentioning green criteria and L-contracts allows for a more accurate attribution of observed effects to environmental considerations.

A "Pre" and "Post" variable was created based on the policy transposition deadline date of April 18, 2016, marking the date by which all EU Member States had to implement the new rules into their national procurement laws. A robust method was implemented to address inconsistencies between dispatch and award dates by always selecting the oldest date and parsing out the corresponding year, month, and day. Additional "announcement period" markers were added to account for the earlier introduction of reforms (February 26, 2014), enabling pre-announcement and post-announcement comparisons.

The dataset was further reduced to relevant analytical variables for treatment-effect estimation, filtering out all redundant or unused fields. The relevant variables will be described in Section 4.4.

Lastly, the first and last appearance dates of both green and non-green (L) contracts per CAE were determined. This allowed the enforcement of a vital constraint: only CAEs with at least one L contract before their first green contract were retained. This ensured straightforward treatment assignment over time.

As a final filtering step, CAEs with only a single appearance in the dataset were excluded, and only those with at least two observations were retained to allow for before/after comparisons.

Final Dataset Description and Readiness

This extensive cleaning and preparation process resulted in a structured and consistent dataset of 18 712 observations and 26 variables, ready to undergo the final operational steps required to implement the research design and assess the associated assumptions. These remaining steps will be outlined in Section 4.4, which delineates the final operational steps needed for analysis.

Data Quality and Limitations of TED

Despite the extensive cleaning, transformation, and filtering efforts detailed above, the used dataset remains shaped by the quality of the underlying source data. As the dataset originates from the TED platform, it is helpful to acknowledge and critically reflect on its structural limitations, as highlighted in existing literature.

TED is the EU's public online platform dedicated to public procurement across EU Member States, registering all tenders exceeding predefined thresholds [6][55][59]. TED falls in line with the EU's initiative to increase transparency and accountability, enabling the distinction between contracts with and without environmental criteria and allowing the tracking of contracts during their life-cycle [55][59].

Prier et al. (2018) [55] highlight that even though TED improves transparency, the data's comprehensibility and structure are challenging for inexperienced users. With data harvested from several standardized forms, the complexity of the structure is very high [55]. Other issues involve data redundancy, duplicate fields, missing values, and the overall lack of standardization [55][59].

The significant presence of missing values within the datasets is deemed a major issue [61]. Vital data fields, such as contract values, duration details, and award information, are often incomplete. Non-compulsory fields are regularly left blank, while mandated fields are sometimes circumvented using placeholder values (e.g., zero) [61]. These deficiencies can hinder research progress, making it essential to address them before any analysis can be conducted. Missing, incomplete, and erroneous data pose a fundamental issue in the European public procurement data system [2]. Between 2009 and 2015, nearly 15% of all mandatory fields were left empty.

However, as Ackermann et al. (2019) [2] point out, the primary limitation to analytical capabilities is not merely the absence of mandatory data. Even when all required fields are completed, the dataset may still lack essential information for meaningful analysis, as critical fields, such as the number of bids or contract value, are not classified as mandatory. Over the years, the percentage of missing values declined, but this does not necessarily dictate an increase in the quality/accuracy of the data [2].

A second issue is the large number of duplicate notices within the database. This arises primarily from legal and reporting frameworks that require contracting authorities to issue new notices at various stages of the procurement life-cycle (e.g., contract notices, awards, amendments) [61]. Instead of updating existing records, a new entry is created for every change or cancellation, leading to redundant records for the same contract [55].

Poorly formed data present challenges, such as non-numerical values in numerical fields and improper use of separators, which complicate data processing. Another concern is erroneous data, with examples including nonsensical timelines (e.g., start dates later than end dates) and abnormally large contract values. Often due to human oversight or inexperience, these errors can negatively impact analyses [61]. Ackermann et al. (2019) [2] highlight instances of erroneously high contract values within the TED database, noting that while only a limited number of such cases have been identified, their presence suggests that many more may exist among the millions of contract award (CA) records from 2009 to 2017.

Another significant issue within TED data relates to framework contracts, which are large, multi-annual agreements between two parties (CAEs or economic operators) that focus on specific thematic areas or methodologies that set the terms for future purchases, allowing public buyers to award contracts without running a full tender each time. Two primary data entry problems arise in TED regarding these contracts. First, some entries attribute the entire framework value to each included economic operator, leading to inflated totals and misrepresentations of contract awards [2]. Second, in other cases, the total framework value is recorded only once, with null values assigned to different operators or entries, distorting subsequent analyses [2]. Furthermore, awarded contracts are often not updated when necessary, compounding inconsistencies and compromising the overall reliability of the dataset [2].

TED data from 2009-2017 is characterized by low quality and reliability due to large amounts of missing data, inconsistency in the way data is entered, and errors in the data entered, including impossibly high figures for the value of some contract awards [2]. Ackermann et al. 2019 [2] note that the gaps and errors in the 2009-2017 data are attributed to the limited number of compulsory fields and the fragmented and complex procurement systems within member states. A lack of compatibility was discovered between different member states' systems and the central TED database, in combination with the need to enter data twice (in both member state systems and TED), and lack of compliance with and enforcement of EU legislation results in limit the analytical capabilities of TED data. Ackermann et al. (2019) [2] highlight that this is logical because, at its core, the TED platform is a publication tool, not designed for analytical purposes.

The limitations of TED's data quality could profoundly impact PP analyses. Results derived from TED must be interpreted cautiously due to data sparsity and quality issues. Approximately 25% of data points are incomplete, with potential misfiling despite efforts by TED personnel to address errors [56].

Soylu et al. (2022) [61] explored the overlap between contract notices and award notices in TED, revealing only a 9% linkage between these datasets. This indicates that only a fraction of announced contracts can be directly matched to their corresponding awards. Such weak data linkages limit the ability to trace contracts from announcement to awarding, hindering insights into procurement efficiency, competition, and transparency [65].

The TED database has also been criticized for requiring extensive data cleaning before analysis [55]. A key takeaway from Prier et al. (2018) [55] is the necessity of a clear strategy to handle the immense number of scenarios that awarding a tender can create to lower complexity during analysis.

Ackermann et al. (2019) [2] identify potential causes for gaps and errors in TED data, attributing these issues to systemic factors. Among these are the limited number of mandatory fields, which can lead to significant omissions in critical procurement information; the fragmentation of national procurement platforms, resulting in inconsistencies in data collection and reporting across Member States; and the complexity of procurement systems, characterized by a multitude of non-standardized templates and forms that hinder harmonization [2]. The lack of compatibility between national procurement platforms and TED has also contributed to double entries and inconsistencies during data transfers [2]. Structural limitations within the TED platform add to the data quality concerns, as the system is designed to fit within an outdated framework that does not prioritize large-scale validation or analytical capabilities [2]. The listed factors diminish the reliability and usability of TED data for comprehensive public procurement analysis.

The outdated standard prompted a major system reform to align with the 2014 procurement directives. The EC aimed to successfully implement the reforms by 2023, marking a 12-year transition since the initial impact assessment of 2011, which highlighted these issues [2].

The reform shifted from a 'paper logic' to an 'IT logic' approach to simplify data entry and transfer. In 2018, national procurement systems and TED were electronically aligned and integrated, hoping to eliminate double entries, reduce gaps, and limit errors. This alignment adheres to the 'only-once' principle to streamline procurement processes [2]. In 2019, the eForms regulation established a single data entry point for automatic data transfers to TED [2]. It also increased the number of compulsory fields and reduced the required forms from twenty-five to six. Moreover, it introduced automatic validation of submitted data, allowing notices to be rejected when they are non-compliant with regulations.

Ackermann et al. (2019) [2] assert that these reforms represent noteworthy progress. However, their success remains reliant upon effective implementation and enforcement. The discretion granted to contracting authorities regarding key data fields continues to raise concerns about the completeness and accuracy of data [2]. Currently, TED's potential as a reliable database for policy analysis remains constrained; in addition to the technical improvements, it is suggested that institutional oversight be increased to ensure rigor, quality, and transparency [2].

4.3 Two-Period Difference-in-Differences: Critical Reflection

After exploring the PSM-DiD and staggered DiD frameworks and acknowledging their theoretical strengths and weaknesses, the final methodological choice was guided by practical constraints related to the structure and quality of the available data. The selected approach resembles the design used in the staggered DiD model, as CAEs are still tracked over time. However, the number of time periods was reduced to two: a consolidated pre-treatment period and a post-treatment period. This simplification effectively addresses the challenges of an unbalanced panel structure, where many CAEs are only observed sporadically in a limited number of years.

The final specification can be described as one-way fixed effects Difference-in-Differences model with CAE fixed effects, working with aggregated panel observations per CAE over two periods (pre- and post-treatment). Due to the heterogeneous and voluntary nature of policy adoption, this CAE level structure enabled more precise identification of treatment and control groups, which was not feasible when attempting to define treatment groups based on broader categories such as country/region, CAE type, or CPV codes. Given the availability of only two periods, the model does not absorb time fixed effects; instead, the post-treatment period is captured explicitly through a binary post indicator, and is therefore not a full two-way fixed effects specification in the classical sense. Yet, this approach offers a robust framework for estimating treatment effects while controlling for time-invariant differences across units and macroeconomic shocks, even under noisy or incomplete data conditions [9].

The current model is classified as a one-way fixed effects design, using a post-period dummy to capture time variation. Nonetheless, it is essential to acknowledge that in future research, where multiple discrete time periods could be used instead of consolidated pre/post periods, the specification would evolve into a generalized two-way fixed effects (TWFE) DiD model. Note that TWFE models have well-documented limitations in settings with heterogeneous treatment timing, as highlighted by Callaway and Sant'Anna (2021) [5]. In scenarios with heterogeneous treatment timing, standard TWFE models may incorrectly compare already treated units with newly treated ones, leading to biased or confusing results [5]. They can also give too much or even negative weight to specific comparisons, making it hard to interpret the overall effect and hiding essential differences between groups [5].

These concerns initially motivated the exploration of a staggered DiD framework in this study. However, the staggered design proved infeasible due to irregular observation patterns and difficulties validating key assumptions, such as parallel trends. Some limitations can be worked around by adopting a simplified two-period setup where treatment does not vary across time within units. According to Callaway and Sant'Anna (2021) [5], TWFE is considered robust in a canonical two-period, two-group setting, closely resembling the structure used here. Accordingly, applying a one-way fixed effects DiD model (with an explicit post-treatment indicator) provides a reliable and appropriate framework for estimating treatment effects in this context.

The final research design was chosen because it is better compatible with the limitations of the dataset. It also performs more robustly under noisy or sparse data conditions than the previously attempted methods. Furthermore, the two-period setup and CAE-level aggregation enhance interpretability. This choice represents a trade-off between methodological correctness and empirical feasibility. While the design is not as sophisticated as the staggered DiD model, it better aligns with the dataset's characteristics, specifically the irregular CAE activity, missing variables, and occasional ambiguity in data entries. Integrating CAE-level fixed effects reduces the risk of unobserved time-invariant heterogeneity across, as it controls for every aspect of a CAE that does not change over time, even if it is unmeasured [31][47]. This makes fixed effects an excellent tool for this study, as many important CAE characteristics (size, experience, complexity of works, procurement capacity, geographic factors, and professionalism) are not captured.

By applying fixed effects, time invariant characteristics of each CAE are removed, ensuring that estimations capture the treatment effect within a CAE over the two periods. In other words only changes within the same CAE (from pre to post) contribute to estimating the impact of adopting environmental criteria. This is called the *within effects* approach, as it compares each CAE to itself over time, rather than between CAEs, thus eliminating bias from any fixed differences [47]. Hence, interpreting estimated effects should be limited to realistic within-CAE changes, instead of deriving results from differences between inherently different entities. Mummolo and Peterson (2018) [47] underline the importance of using within-effects by recognizing that within-unit variation is always smaller or equal to the overall variation in the dataset. Therefore extrapolating treatment effects to differences across different CAEs, particularly when they involve larger changes than any observed within a CAE, would lead to extreme, unrealistic counterfactuals [47].

By focusing on the within-unit change, the current model strengthens the causal inference under the DiD configurations.

To better explain what fixed effects do, consider the following scenario: Suppose CAE X is a large national authority that generally tenders major infrastructure projects, attracting an average of four bids per tender due to the limited number of firms capable of handling these complex projects. CAE Y, a smaller regional authority mainly tendering for minor maintenance works, has a larger supplier base, receiving an average of ten bids per tender.

Without fixed effects, the model could attribute the lower number of bids at CAE X to the usage of green criteria, even if the true reason is that the project complexity and scale limit the pool of eligible suppliers. With fixed effects, the model absorbs this complexity aspect, allowing the analysis to isolate whether introducing environmental criteria is associated with an increase or decrease of bids for each CAE individually relative to its pre-treatment levels.

Despite its advantages, the approach still relies on strong assumptions. One key decision involved defining and operationalizing a clear and consistent scope of the treatment. The 2014 Public Procurement Directive encouraged CAEs to move beyond lowest-price evaluations by allowing the inclusion of environmental, social, and innovation-related criteria. While the directive broadly promotes using non-price criteria, this study specifically focuses on estimating the effect of explicit ecological considerations in public tenders.

To support this more narrowly defined treatment, all M contracts (i.e., those using MEAT as the award criterion) that did not include environmental criteria were excluded from the dataset. This exclusion helps construct a cleaner comparison by reducing noise, because including non-green M contracts in the control group could confound the results, as other non-price criteria might also influence tender outcomes, such as cost and competition. The resulting setup allows for a more targeted estimation of the effect of environmental criteria in EU construction procurement.

However, several limitations must be acknowledged. First, removing non-green M contracts does not exclude the associated CAEs from the dataset. As a result, CAEs that previously used other non-price criteria may still exhibit altered behavior in subsequent L contracts, either due to organizational learning or because prior use of MEAT criteria may have influenced how suppliers approach future tenders. This residual impact cannot be fully accounted for in the current design, but the adopted approach still represents a reasonable approximation of the treatment effect.

Second, as Testa et al. (2016) [63] highlight, environmental criteria can be applied across five procurement components: subject matter, technical specifications, selection criteria, award criteria, and contract performance clauses. However, the dataset only captures award criteria recorded in the CRIT_CRITERIA field. As such, this study can only identify a contract as "green" if environmental criteria were included in the award phase, correctly reported in TED, and matched to a predefined list of environmental keywords and associated terms. This implies that some contracts that applied green criteria in other components may be misclassified as non-green, thereby underestimating the presence of green procurement.

Another central assumption is the parallel trends assumption, which states that treated and control units would have followed similar outcome trajectories over time without treatment. Since adoption is not explicitly staggered in this simplified model, variation in treatment timing is not captured, limiting the ability to assess heterogeneous effects over time.

Another drawback of a two-period DiD framework is that the assumption of parallel trends cannot be formally tested, as a trend requires at least two pre-treatment periods. There is no observable trend with only one pre-period, and the assumption must be justified conceptually rather than statistically.

However, preliminary analyses could be conducted using the unconsolidated dataset to strengthen confidence in its plausibility, where yearly outcome patterns across treatment and control groups were visually compared, and interaction regressions were performed using pre-treatment data. After consolidation, direct year-by-year trend analysis was no longer possible. The methodological steps and diagnostics applied are described in Section 4.4, detailing how the assumptions were examined using visual diagnostics, statistical tests, and robustness checks.

There is also a risk of omitted variable or selection bias, as the model only controls for covariates observed in the dataset. Several potentially essential variables, such as project duration, evaluation weightings, or justification for award decisions, were unavailable or unusable due to formatting inconsistencies or data quality issues. These gaps may influence treatment assignment and procurement outcomes in ways that cannot be directly controlled. While the one-way fixed effects Difference-in-Differences model accounts for time-invariant unobserved heterogeneity across units, it does not address time-varying or tender-specific unobservables [16].

This study would have benefited from a more complete and balanced dataset where each CAE could be consistently tracked across all years. This would have made it easier to compare changes over time and strengthen the overall reliability of the analysis. Access to more detailed and reliable project-level information, such as precise data on project duration, awarded amounts, and the specific evaluation process, would also have improved the quality of the results. Additionally, it would have been useful to have a way to deal with the possibility that green tenders are selected for reasons not visible in the data, such as internal sustainability strategies or political priorities. This could have been addressed by adding extra variables or external data sources to help separate genuine treatment effects from background influences. Finally, a more standardized set of correctly inputted variables could have supported stronger comparisons between tenders.

Unfortunately, such enhancements were not feasible within this thesis's current dataset and time constraints. Given these limitations, the selected approach offers a sound and practical approach for estimating the causal impact of environmental criteria on procurement outcomes in the EU construction sector.

4.4 Operational Implementation

After a comprehensive data cleaning and preparation process, it was determined, given this study's scope and time constraints, that a one-way fixed effects Difference-in-Differences analysis would offer the most appropriate methodological fit. However, several operational steps were required to translate the theoretical framework into a practical analytical format.

The following paragraphs outline how the key outcome variables, project cost and bidding competitiveness, were defined based on the feasibility and relevance of the available data. Next, the rationale behind the tested configurations is briefly introduced. These differ in how consistently treated behavior must be observed after the policy change, ranging from a stricter setup, where green criteria must be used continuously, to more flexible interpretations that allow occasional reversion to non-green tenders. This variation helps assess the robustness of results under different assumptions about treatment persistence.

Subsequently, the essential control variables used in treatment-effect estimation are explained, including the logic behind the consolidation process (e.g., use of means, modes, or first observations). The degree of correlation between these covariates is also briefly assessed to evaluate potential multicollinearity concerns. Along with the justification behind the usage of clustered standard errors.

Finally, the assumption of parallel trends is examined using visual plots and interaction regressions on pre-treatment data. These results also justify using the policy announcement date (26 February 2014) as the cut-off between pre- and post-treatment periods. A short explanation of how the counterfactual was calculated to support the interpretation of the DiD results.

RStudio was used to perform this analysis. It is an integrated development environment for R, a statistical computing and graphics programming language. R was selected for its flexibility and wide range of statistical packages to support efficient data manipulation and visualization. The dataset required extensive restructuring, filtering, and aggregation, which were efficiently executed using tools like *dplyr*. For modeling, *fixest* allowed the implementation of one-way fixed effects regressions with clustered standard errors.

Operationalization of Outcome Variables

The contract award value, in euros without VAT, was selected as the best approximation of the project cost. As previously mentioned, steps were taken to ensure consistency. First, all monetary values were converted to constant 2023 euros. Multiple value fields were compared to create a single final_price_contract variable. Then, for each call for tenders, the total contract value was calculated by summing the values of all associated contract awards. To address data quality concerns, extreme outliers were filtered out based on a visual inspection. Given the substantial right-skewness of the distribution, a log transformation was applied to normalize the variable and improve interpretability in regression models. This transformation helps stabilize variance and ensures that large contracts do not disproportionately drive the results. The outcome metric for each CAE was computed as the mean log contract value for the pre-treatment and post-treatment periods, respectively.

For this analysis, given the available data's size and limited explanatory capacity, adopting the number of bidders as a proxy for the level of competition seems most feasible. Focusing the analysis solely on the Single Bid contracts was found to be an insufficient analytical approach by Tátrai et al. (2024) [65]. The total number of offers received across all awarded CAs within a tender call was used to operationalize this, providing a more continuous and informative indicator of competition intensity. Similar to contract price, the distribution of bids was right-skewed, prompting a log transformation. After filtering for extreme values, the average number of offers each CAE received per call sent out was calculated for the pre- and post-treatment periods.

Visual illustrations of the distributions before and after natural log transformation can be found in Appendix 7.5. For both variables of interest, the distribution (Fig. 5.1 and 5.5) and the Q-Q plot (Fig. 5.3 and Fig. 5.7) indicated that the values were heavily right-skewed, exhibiting substantial tail weight and observable divergence from normality. Applying a natural logarithmic transformation improved the alignment with the normal distribution and symmetry, especially in the central quantiles (Fig. 5.2, 5.4, 5.6 and 5.8). At extremes minor deviations persisted, nonetheless the transformation reduced the impact of outliers and decreased variance.

Overview of Tested Configurations

Before introducing the three tested configurations, a series of core filtering steps was applied to create a clean and comparable dataset that was the foundation for all subsequent setups. These steps aimed to retain only contracting authorities (CAEs) with meaningful pre- and post-treatment activity.

The first step was to exclude any CAEs whose first green contract occurred before the official announcement of the 2014 Directive (26 February 2014)[8]. Only units that could reasonably serve as untreated comparisons in the pre-period were retained. This step reduced the dataset from 18 712 to 18 320 observations.

Second, CAEs were required to have all pre-treatment contracts use the lowest-price evaluation and have at least 2 calls for tenders across both periods. This reduced the dataset to 10 395 observations.

Third, CAEs that appeared in only one year were excluded, as they could not contribute to the panel structure needed for longitudinal analysis. This step removed 135 CAEs, accounting for 348 observations, resulting in 10 047 remaining observations.

Finally, only CAEs observed in both periods were retained. This step ensured a proper DiD setup and reduced the dataset to 915 CAEs and 9 785 observations. A total of 262 observations belonging to 80 CAEs that were only present in a single period were excluded at this stage.

Following these steps, CAEs were classified as treated if they had used environmental award criteria in any post-treatment contract, and control otherwise. This classification resulted in 665 CAEs in the control group and 250 CAEs in the treatment group.

It was necessary to consolidate all variables that could attain multiple values for each observation of a CAE within these periods in order to facilitate the two-period DiD design with a single pre- and post-treatment observation per CAE. This consolidation was carried out separately for each tested configuration, after the configuration inclusion logic had been applied.

For numerical variables such as project cost and the number of offers, the mean was computed within each period (pre or post). Categorical variables were collapsed to the most frequently observed value. In case of ties, the first encountered value was retained. The corresponding code can be found in the Appendix 7.1.

To test the robustness of the estimated treatment effects, three alternative treatment definitions were applied, each with its own inclusion logic and assumptions about behavioral persistence following the first adoption of green criteria. These definitions led to different sample sizes, as summarized in Table 2.

- **Only green in post period:** In this strict configuration, only CAEs whose post-treatment contracts all used green criteria were retained in the treatment group. Any CAE that used an L contract after the announcement of the directive was excluded. This led to removing 162 CAEs and 3 717 observations, leaving 753 CAEs (88 treated and 665 control) and 6 068 observations.
- **Once treated, remain treated:** This setup retained CAEs that, from the point of first adopting green criteria onward, never returned to lowest-price-only tenders. To account for a clean transition, treatment was defined as consistent green behavior after the first instance of adoption, and only CAEs meeting this rule or remaining fully untreated were retained. This resulted in 819

[8] The logic behind this cut-off is discussed later

CAEs (154 treated and 665 control) and 6 723 observations. A total of 3 062 observations from 96 CAEs were excluded.

- **Treated once is enough:** In this most inclusive configuration, a CAE was classified as treated if it adopted green criteria at least once after the policy announcement. No further constraints were imposed on later behavior. This setup preserved the full filtered sample of 915 CAEs (250 treated and 665 control), resulting in 9 785 observations.

Assessment of Multicollinearity and Covariate Independence

Each configuration was estimated using the same set of covariates to ensure comparability. While fixed effects account for time-invariant characteristics of each contracting authority entity (CAE) [31][47], these covariates were included to control for time-varying contextual factors that may influence outcomes independently of the treatment. The regressions included an *outcome* variable, either the log-transformed total project cost or the log-transformed number of offers received. The variable *post* is a binary indicator for the post-treatment period, while *treat* marks whether a CAE is in the treatment group. *SDG* captures the Sustainable Development Goal index score for the CAE's country, reflecting national environmental ambition. *GOVexp* represents public expenditure as a percentage of GDP, a proxy for government investment capacity. *Spillover* accounts for the international environmental and social effects a country generates through trade and finance, which may influence procurement expectations. *Covid* reflects pandemic intensity at the time of tendering, scaled per country and month. *CPPI_growth* measures the average annual growth rate of the Construction Producer Price Index, controlling for sector-wide pressures that could impact project costs and bidding behavior. Finally, *CAE_TYPE_grouped* captures high-level institutional types (e.g., national, sectoral), while *region* or *country* adjusts for geographic variation without relying on individual countries with limited sample sizes.

Note that, since CAE-level fixed effects already capture all time-invariant differences across contracting authorities, including *CAE_TYPE_grouped*, *region*, and *country* directly would be redundant. However, they are retained in the simpler linear models without fixed effects.

Table 2: Sample sizes and treatment classifications under different assumptions

Overview of Tested Treatment Definitions			
Configuration Description	**Total CAEs**	**Treatment CAEs**	**Observations**
Only green in post	753	88	6 068
No reversal after first use	819	154	6 723
Treated once is enough	915	250	9 785

Before interpreting the regression results, it is vital to check whether any of the control variables are too closely related to each other. If two or more variables are strongly related, it can make it harder for the model to estimate their individual effects clearly. Two common diagnostic tools were used to examine this: pairwise correlations and Generalized Variance Inflation Factors (GVIFs). The results are shown in Table 3 and Table 10.

Table 3 shows how strongly each pair of continuous variables is related. Most relationships are weak, which is a good sign because it means the variables are not overlapping too much. One exception is the correlation between the SDG score and GOVexp. This relationship is quite strong across all model configurations, with values between 0.4773 and 0.4898. This may reflect a tendency for

Table 3: Pairwise correlations between key covariates across tested configurations

Variable Pair	*Only Green*	*No Reversal*	*Only Once*
cor(SDG, Spillover)	-0.0649	-0.0846	-0.0801
cor(SDG, GOVexp)	0.4773	0.4808	0.4898
cor(Spillover, GOVexp)	-0.4191	-0.4291	-0.4197
cor(COVID, Spillover)	-0.1033	-0.1037	-0.1143
cor(SDG, COVID)	0.0840	0.0873	0.0836
cor(COVID, GOVexp)	-0.0355	-0.0363	-0.0416
cor(CPPI_growth, Spillover)	0.1112	0.1065	0.1055
cor(CPPI_growth, SDG)	0.0813	0.0904	0.0938
cor(CPPI_growth, GOVexp)	0.2457	-0.2424	-0.2441
cor(CPPI_growth, COVID)	0.1969	0.2002	0.2114

countries with higher sustainability performance to also exhibit higher levels of public sector spending.

There is also a moderate negative relationship between spillover effects and GOVexp (between -0.04191 and -0.4291). This suggests that high public spending doesn't necessarily lead to more positive international effects. Meanwhile, the relationship between SDG scores and spillover scores is weakly negative (around -0.07), indicating that progress at home doesn't always translate into a positive influence abroad. All other relationships, especially those involving the COVID variable, are closer to zero. This is good because it means these variables will likely capture different information and not duplicate each other.

To analyse GVIF, a linear regression model (Ordinary Least Squares) was run to analyze the determinants of our outcome variables. The results of this model are presented in Table 9 in Appendix 7.7, but are not further discussed. Given the study's causal objective, OLS would inadequately address heterogeneity and would implicitly assume cross-sectional comparability between CAEs, a condition unlikely to hold in this context.

In the final one-way fixed effects DiD model, fixed effects at the CAE level are included, which automatically absorbs time-invariant features [31]. To assess potential multicollinearity among the control variables, an initial linear regression was estimated without fixed effects. Fixed effects absorb part of the variation in explanatory variables, altering how relationships between them are measured and potentially masking collinearity [39]. By first estimating a simple linear regression without fixed effects, multicollinearity among the relevant control variables can be more easily detected. The OLS model specification is:

$$\text{Outcome} = \beta_0 + \beta_1 \text{ post} + \beta_2 \text{ treat} + \beta_3 (\text{post} \times \text{treat}) + \beta_4 \text{ SDG} + \beta_5 \text{ GOVexp} + \beta_6 \text{ covid} \quad (1)$$

$$+ \beta_7 \text{ construction\_growth } (+ \beta_8 \text{spillover}) + \sum_j \delta_j \text{CAE\_TYPE\_grouped}_j + \sum_k \theta_k \text{COUNTRY}_k + \epsilon_{it}$$

However, checking for multicollinearity in the OLS model has limitations, as it reflects overall correlations rather than within-group variation. Correlations may disappear or appear after applying fixed effects. Thus, while this check offers a useful initial screening, it does not fully capture collinearity in the final model.

Table 10 in Appendix 7.8 presents the GVIF analysis results. This also helps to detect multicollinearity, which happens when variables in a model are too closely related. High multicollinearity can make the results unstable or unreliable. GVIF is a version of the standard Variance Inflation Factor (VIF) that works better when you have categorical variables with multiple categories (like country or CAE type). Because some variables have more than one degree of freedom (Df), their GVIFs are adjusted using the formula:

$$\text{GVIF}^{1/(2\times \text{Df})}$$

This adjustment makes the GVIFs easier to compare to the usual thresholds (values above 10 are seen as problematic, between 5 and 10 caution is advised). The table presents all the results

Note that The variable *spillover* is shown between parentheses because the model was estimated twice: once including, and once excluding this variable. Due to the consistently high Adjusted GVIF values associated with *spillover*, around 12 across all model configurations, it was ultimately omitted from the main specification because of clear evidence of multicollinearity. The GVIF values presented correspond to the model without *spillover*. However, the value for *spillover* is still reported to justify its exclusion.

Looking at the adjusted GVIFs, SDG and GOVexp again show the highest values (larger than 5 across models). These are noteworthy and confirm that the two variables are somewhat related, just as the correlation analysis suggested. Both variables are retained, given their expected importance, but results must be interpreted cautiously. No meaningful multicollinearity is detected among the other controls.

The country variable had very high unadjusted GVIFs (over 12), but its adjusted values remain around 1.17. This is reassuring and means that the model handles this complexity well.

Based on the correlations and GVIF values, only spillover showed problematic multicollinearity. While a few variables, especially SDG and GOVexp, are moderately related, this is not strong enough to harm the model. Therefore, all other variables were kept in the final analysis to ensure the model controls for essential differences between CAEs and countries.

Clustered Standard Errors

Given the panel structure of the dataset, residuals could have potentially complex dependencies with each other. Such dependencies lead to biased standard errors, therefore invalidating the statistical significance of the estimated coefficients [31].

To account for potential intra-cluster correlation in the residuals, standard errors (SE) were clustered [31]. The idea is that by clustering, the residuals can have arbitrarily shaped forms and be dependent instead of being considered independent, which is not realistic when working with panel data. Clustering SE allows the residuals to be dependent within each cluster, while still assuming independence across them.

Several clustering options were considered, including clustering at the region (n=5), country (n=27), and CAE (n>750) levels. Clustering at the CAE level was ultimately chosen, as it aligns with the study's treatment assignment framework. Treatment variation occurs at the level of the entities, therefore clustering at broader levels, such as country or region, would obscure important within-country and within-region differences, as contracting authorities within the same geographic entity could have different procurement aims. Ignoring this variation would misrepresent the actual structure of the residuals, potentially underestimating intra-group correlation patterns [31].

Moreover, clustering at the CAE level benefits from a sufficiently large number of clusters, supporting robust inference. A best practice for clustering SE is having a sufficiently large number of clusters, because the consistency of the cluster-robust SE is positively correlated with the number of clusters [31].

In this setting the treatment originates from the EU 2014 PP directive encouraging the use of environmental criteria in public procurement. While countries had discretion over the timing and manner of transposing this directive into national law (deadline: 18/04/2016), the decision to apply environmental criteria was made individually by contracting authorities and independent of national legislation. Consequently, treatment assignment varies at the CAE level, further justifying the decision to cluster standard errors at this level.

Parallel Trends Assumption

The parallel trends assumption is central to the internal validity of a DiD-analysis and arguably the most difficult to satisfy. It requires that the average outcome trends for treated and control groups would have remained parallel over time, if treatment had not occured [7]. To substantiate this claim the trendlines prior to treatment are compared.

While the assumption cannot be statistically tested in a standard two-period setup, it must be validated conceptually since a trend requires at least two pre-treatment periods. This study's two periods are synthetic: each CAE's pre-treatment period is constructed as an average across multiple years. This means preliminary analysis using the unconsolidated dataset to reinforce the plausibility of parallel trends is possible. Specifically, yearly outcome patterns were visually inspected, and interaction regressions were estimated using pre-treatment data. These regressions at different moments also serve as placebo tests, as they estimate treatment effects using hypothetical cut-off dates [40]. The absence of significant effects at these placebo points could provide additional support for the assumption.

This section presents the applied tests and their results, justifying the policy announcement date (26 February 2014) as the cut-off between pre- and post-treatment periods. Each configuration was individually tested. While the control units remained constant across setups, the treatment groups varied in both size and composition. In addition to these structural differences, extensive testing was conducted to determine the most accurate and robust cut-off date for identifying the treatment effect. This was based on a combination of visual inspection of yearly trends, joint F-tests of pre-treatment interactions, and logical reasoning related to the policy timeline. For both outcome variables, line plots were generated to compare pre-treatment trends across all configurations and under multiple candidate cut-off dates. Shifting the cut-off date also impacted group sizes, influencing the treatment-control dynamics.

Prior to analysis, six potential cut-off dates to distinguish the pre- and post-treatment periods were identified and evaluated: January 1, 2014; the policy announcement date (February 26, 2014); January 1, 2015; January 1, 2016; the formal transposition deadline (April 18, 2016); and January 1, 2017.

A combination of logical reasoning, visual inspection of outcome trends (Appendix 7.3), and statistical testing (Appendix 7.2) was employed to identify the most appropriate date. These analyses were performed using the unconsolidated dataset version, allowing multiple years of pre-treatment observations. To estimate the interaction regressions, the following model was specified:

$$Y_{it} = \alpha + \sum_k \beta_k \cdot \text{Year}_k + \gamma \cdot \text{Treat}_i + \sum_k \delta_k \cdot (\text{Year}_k \times \text{Treat}_i) + \varepsilon_{it} \tag{2}$$

Here, Y_{it} denotes the outcome variable for unit i in year t. Treat_i is a binary indicator for treatment group membership, and Year_k are year dummies for pre-treatment years, with one base year (2012) omitted. The interaction terms $(\text{Year}_k \times \text{Treat}_i)$ capture whether the treated and control groups followed different trends in those years before treatment. The coefficients δ_k reflect these differences.

A joint F-test was used to test the null hypothesis that all δ_k equal zero. Failure to reject this hypothesis indicates no evidence of statistically diverging pre-treatment trends between treatment and control groups, thus supporting the assumption of parallel trends.

Note that no covariates were included in these regressions. Since the assumption of parallel trends pertains to the evolution of the outcome variable across groups, adjusting for covariates at this stage could mask genuine deviations and lead to incorrect confirmation of the assumption. In the Difference-in-Differences model, covariates will be included later to improve precision, but are excluded from the assumption check.

The joint F-tests for contract values and received tenders, summarized in Table 5 and Table 6 respectively, provide an initial statistical basis to evaluate the plausibility of the parallel trends assumption across all candidate cut-off dates and configurations.

In Table 5, no joint p-values fall below the 5% significance threshold across any configuration or cut-off. A few individual interaction terms, notably the "Only Green" configuration at the 2015-01-01 cut-off and "No Reversal" at the 2017-01-01 cut-off, have individual p-values below 0.10, which are suggestive but statistically weak. These results are insufficient to reject the assumption of parallel trends in those cases. In contrast, Table 6, which covers the tests for received tenders, shows clearer violations. In particular, two cut-off dates can be statistically ruled out:

April 18, 2016 (Transposition Deadline):

For the "Only Green" configuration, the joint p-value is 0.0487, indicating rejection of the null hypothesis of equal trends at the 5% level. Additionally, one interaction term (2014:treat) is significant at the 5% level, suggesting divergence in pre-treatment trends. The "Only Once" configuration similarly shows a joint p-value of 0.0312, reinforcing rejection of the parallel trends assumption at the 5% level. Within this group, the 2016:treat interaction term is highly significant ($p < 0.01$), further highlighting substantial deviations. The "No Reversal" configuration also shows some evidence of divergence, with the 2016:treat term significant at the 10% level.

These results imply that the treated and control units did not follow parallel pre-trends under this cut-off. Visual inspection of the plots 3.9 (Contract Value) and 3.10 (Received Tenders) confirms these statistical findings.

In Fig. 3.10, the "Only Green" group initially converges with the control group until 2014, but then diverges again in 2015–2016. While the "No Reversal" and "Only Once" groups appear to move roughly parallel to the control up to 2014, they exhibit greater volatility in trend from 2014 onwards. In Fig. 3.9, although the absolute differences in mean contract values across groups are moderate, the direction of trends is inconsistent, especially between 2013 and 2015, rendering the assumption of parallel pre-treatment trends visually questionable.

From a logical reasoning perspective, this date is also problematic. CAEs were not legally required to wait until national transposition before implementing the 2014 directive. In practice, they could adopt environmental criteria once the directive was announced, unless explicitly restricted by national procurement laws, which was unlikely.

As such, selecting the transposition deadline as the cut-off risks contaminating the pre-treatment period, since treated CAEs may already have adapted their behavior. This would violate the fundamental premise of DiD designs and introduce bias into the estimation. Therefore, based on statistical evidence, visual inspection, and logical argumentation, the transposition deadline (April 18, 2016) was deemed unsuitable as a treatment cut-off for this analysis.

January 1, 2017:

The "Only Once" configuration shows a highly significant joint p-value of 0.0087, with three individual years (2013[9], 2014[10] and 2016[11]) showing significant p-values. This suggests substantial divergence between the treatment and control trends before treatment within this configuration. The "No Rever-

[9] $p < 0.01$
[10] $p < 0.1$
[11] $p < 0.05$

sal" configuration also violates the parallel trends assumption. The joint p-value is 0.0309, indicating rejection of the null hypothesis of equal trends at the 5% level.

The corresponding plots (Fig. 3.11 and Fig. 3.12 support this, revealing diverging and volatile trends, especially after 2014 for the received tenders.

Logically, selecting this date as the cut-off would not align with the timeline of how the policy was announced and implemented. By the first of January 2017, contracting authorities could already have adopted green criteria in practice for several years. Including such observations in the pre-treatment period risks contaminating it with treated behavior, thereby masking true treatment effects and biasing the results. For this reason, and supported by both statistical evidence and visual trends, this cut-off was rejected.

The remaining candidate dates did not show sufficient statistical evidence to reject the assumption. As shown in the tables in Appendix 7.2, none of the joint p-values for these dates fall below the conventional 10% significance threshold, and most interaction terms are individually non-significant.

However, the absence of statistical significance does not automatically confirm the validity of the parallel trends assumption within this context. Therefore, additional evaluation through visual inspection of the pre-treatment trends and logical reasoning about the directive's timeline is required. In what follows, each remaining candidate cut-off is assessed, focusing on the plausibility of parallel evolution between treatment and control groups and the coherence of the cut-off date within the policy implementation context.

January 1, 2014, 2015, and 2016:

Starting with January 1, 2015 and January 1, 2016, the key concern is their placement after the policy announcement (February 26, 2014). As the announcement marks the first moment that contracting authorities could be exposed to or act upon the policy, these later dates risk classifying already treated units as part of the pre-treatment group. This may bias the analysis and underestimate the estimated effect of treatment, particularly if the treatment is hypothesized to reduce the number of offers and increase contract prices. Visual plots support this concern, showing increasingly diverging or unstable pre-treatment trends as the cut-off moves further beyond the announcement date.

In contrast, the January 1, 2014 cut-off precedes the policy announcement, which avoids the risk of including treated observations in the pre-treatment group. Visually, the trends between treatment and control groups appear parallel and stable. However, this early cut-off also presents limitations. Untreated units from 2014-1-1 until 2014-02-26 would be classified as post-treatment, potentially leading to an overestimation of the treatment effect, as these are hypothesized to be lower than treated units.

Therefore, despite statistical plausibility, none of these dates present a coherent foundation for identifying the treatment onset. Based on these considerations, they are not selected as the final cut-off date.

February 26, 2014 (Announcement):

The policy announcement date, February 26, 2014, was ultimately selected as the preferred cut-off. This moment marks the first official communication regarding the directive and therefore represents a realistic point at which CAEs may begin to adapt their behaviour. Although the interaction term 2014 in the "Only Once" configuration for the offers outcome is statistically significant at the 5% level ($p = 0.0377$), this deviation is interpreted cautiously. The number of observations before the cut-off in 2014 is limited to approximately three months, reducing the reliability of year-specific effects for that period. Importantly, the joint F-tests for all configurations yield non-significant results, and visual inspection of the trends (see Fig. 3.3 and 3.4) confirms parallel trends between treatment and control groups during the core pre-treatment years 2012–2013. The divergence observed in 2014 is thus likely due to the short window rather than a genuine violation of the assumption of parallel trends.

Taken together, both the statistical tests and logical reasoning support the announcement date as a valid and defensible cut-off for identifying treatment effects.

Yet, Huntington-Klein (2021) [40] emphasizes the importance of carefully considering how the outcome variable is measured. The parallel trends assumption concerns not only the direction of trends but also how the difference between treatment and control groups is defined. This becomes especially relevant when applying a log transformation: a trend that reflects a consistent absolute difference over time may not reflect a consistent percentage difference after transformation, and vice versa [40]. In this study, the outcome variables are log-transformed, meaning that the parallel trends assumption is evaluated in terms of relative (percentage-based) changes rather than constant numerical gaps. As a result, the analysis focuses on proportional differences between groups, and no claims can be made about absolute treatment effects in the original scale.

Counterfactual Estimation and ATT Interpretation

In a DiD analysis, a major challenge is estimating the unobserved counterfactual outcome: what would have happened to the treated group in the post-treatment period had they not received the treatment [7]. Because we cannot observe this directly, we rely on the control group to proxy this hypothetical scenario under the assumption of parallel trends.

The unobserved counterfactual isolates the treatment effect from confounding time trends or baseline differences. By comparing the actual post-treatment outcome for the treated group with this counterfactual, we obtain the average treatment effect on the treated (ATT).

To visualize this logic, simplified Average Treatment Effects on the Treated (ATT) were calculated using only group-level averages, providing a transparent estimate of policy impact. These values correspond to the estimated effects from the baseline models in Tables 7.9 and 7.11, provided in the appendices. Using these values, the counterfactual outcome paths can be calculated and illustrated in the graphical representations. The ATT approximations can be calculated in two algebraically equivalent ways[12]:

$$\text{ATT} = (\bar{Y}_{\text{treat, post}} - \bar{Y}_{\text{treat, pre}}) - (\bar{Y}_{\text{control, post}} - \bar{Y}_{\text{control, pre}}) \tag{3}$$

$$\text{ATT} = (\bar{Y}_{\text{treat, post}} - \bar{Y}_{\text{control, post}}) - (\bar{Y}_{\text{treat, pre}} - \bar{Y}_{\text{control, pre}}) \tag{4}$$

The ATT values were used to plot a counterfactual trend for the treated group, assuming they would have experienced the exact change over time as the control group. Figure 4 in Appendix 7.4 contains the plots that visually highlight the difference between the observed post-treatment outcome and the estimated unobserved counterfactual, which corresponds to the DiD estimate.

The Average Treatment Effect on the Treated (ATT) is equal to the Difference-in-Differences (DiD) estimate only under certain conditions. Most importantly, the DiD estimator identifies the ATT when the parallel trends assumption holds and treatment timing is consistent. In a standard two-period DiD design like the one used in this study, with a single treatment point and no staggered adoption, the DiD estimate corresponds to the ATT for each configuration. However, if treatment effects are heterogeneous, meaning they differ across treated units, the ATT reflects an average of these varying effects rather than a uniform impact. As further discussed in Section 5, the estimated treatment effect changes when covariates are included, suggesting that the ATT in this study captures heterogeneity driven by observable characteristics.

[12] Here, $\bar{Y}.$ denotes the average of the outcome variable for a given group and period.

It is important to note that these plots visualize the estimated effect of treatment for the baseline models, but do not indicate statistical significance, and are purely for illustrative purposes.

The table below shows the calculated ATT values for each configuration and each outcome variable:

Outcome Variable	Only Green	No Reversal	Only Once
Contract Value	0.162	0.078	0.069
Tenders Received	-0.233	-0.215	-0.164

Table 4: ATT values used for counterfactual plots (based on group averages)

5 Results

In this section of the thesis, the results of the empirical analyses are presented and interpreted. The study can be divided into three parts, each focusing on one of the research questions defined in Section 2:

- **5.1 Utilization of environmental award criteria**: This subsection discusses the distribution and occurrence of environmental award criteria within public construction procurement contracts over different time periods.
- **5.2 Impact of environmental award criteria integration on bidding competitiveness**: This subsection focuses on how integrating green criteria influences bidding competitiveness, proxied by the number of tenders received, within public construction procurement contracts.
- **5.3 Impact of environmental award criteria integration on project cost**: This subsection focuses on how integrating green criteria within public construction procurement contracts impacts project cost, approximated by the projected contract values.

5.1 Utilization of Environmental Award Criteria

Before the 2014 PP directive announcement, CAEs could already award contracts based on non-price criteria. The 2014 directive builds upon the 2004 PP directive, which stated that contracts could be awarded based on the most economically advantageous tender or the lowest price [29]. This was confirmed during the analysis of the prevalence of green criteria in the construction sector, with a considerable proportion of green contracts identified before the announcement date of the 2014 directive.

This observation calls for a critical reflection on what constitutes the actual treatment of the 2014 PP directive. As non-price criteria existed under the 2004 directive, the 2014 directive can be seen less as a first-time intervention and more as an extension, providing increased legal support and clarity. To assess if the strengthened policy signal, did indeed result in a higher utilization grade, the annual usage of environmental criteria was analyzed and compared. This provided valuable insight into how the 2014 Public Procurement (PP) directive impacted the utilization of environmental constraints in tendering processes by comparing proportions over time.

Under the assumptions of this thesis, the treatment group consists of CAEs that did not include environmental criteria in their tenders prior to the 2014 directive announcement but began doing so afterward. As such, the DiD estimates capture the effect of adopting green criteria, prompted or reinforced by the directive, on procurement outcomes, specifically among CAEs that had not previously (between 2012 and 2014) engaged in green tendering. The estimates therefore do not reflect a potential boosting effect on CAEs that were already using environmental criteria before the directive, but instead isolate the impact of first-time adoption.

An earlier version of the dataset, one before filtering based on the first appearances of green contracts, was used to accommodate this analysis. A total of 27 659 answered call for tenders from 9 606 CAEs were analyzed, spanning the period from 2012 to 2023. The relative number of contracts using at least one environmental award criterion was calculated for each period. Section 4 mentioned that these criteria (list provided in Appendix 7.15) were identified using keyword matching and manual corrections of the provided award criterion descriptions.

In Appendix 7.6, two tables are provided, which are the product of the analysis. Table 7 displays the relative share of green contracts over several time-frames. The column *Total CFT* shows the total number of calls for tenders, and *% Green* signifies the proportion of the contracts that included environmental award criteria. *Ex ante* and *ex post* relate to the before and after of the announcement and transposition deadline of the directive in 2014 and 2016 respectively. The *During* period encapsulates the observations between 26/02/2014 and 18/04/2016.

To properly evaluate the impact the directive had on the degree of utilization, proportions comparison tests were used to statistically assess the differences between the proportions of green contracts over different periods. Specifically, the *prop.test()* function in R was applied, which performs a chi-squared test of equal proportions.

For this test, the null hypothesis states that no significant difference exists between the proportions of green contracts across the different periods being compared. Suppose the p-value is lower than the chosen significance level. In that case, there is a statistical contrast in the proportions of green contracts between the two periods.

The following line of R code was used to carry out the test:

```
prop.test(x = c(X1, X2), n = c(N1, N2), alternative = "two.sided", conf.level = 0.95)
```

Where `x` represents the number of green contracts in each group and `n` represents the sample size of each group. The first input corresponds with the value of the earliest chronological period. The output of the test provides the *Chi-squared statistic* (χ^2), the *p-value*, and the *confidence interval*, which provide evidence on whether there is a significant difference in the proportions of green contracts. The results of this test are shown in Table 8.

The analysis shows that the proportions between the *Pre* and *During* periods are not statistically different. The percentage of green contracts shows a clear increase between 2014 and 2017 and in the later years. Before the introduction of the policy measure, only 14.4% of the calls for tenders contained green criteria, this gradually increased, with 31.33% of the tenders in construction being awarded on the basis of environmental award criteria in 2023.

It is remarkable that a strong increase is visible in 2017, where the percentage increases with around 7% compared to 2016. Looking more closely at 2016, the utilization significantly increases after the transposition deadline. This suggests that the 2014 directive did have an influence in promoting the use of environmental award criteria, but that the actual effect only manifested itself several years after the introduction. Potentially not coming into full effect until all member states changed their legislation.

This is confirmed by the χ^2 values and associated p-values of the comparison between the *Pre* and *During* periods. The p-value of 0.9889 shows that there is no statistical difference in the proportions between the two periods. This indicates that there was little change in the use of environmental criteria before 18/04/2016.

A possible explanation for this observed delayed effect lies in the length of the tendering process itself. As discussed in Section 3.3, the full cycle of an international open tender for works contracts typically spans at least 7 to 8 months[13], from initial publication to contract award. Therefore, tenders initiated in 2018, can appear in the dataset as awarded contracts in 2019 or later. In addition, there were known input issues related to the dispatch dates of the contract notices and awards. Specifically, inconsistencies in the DT_Dispatch (dispatch date of the notice) and DT_Award (contract award dispatch date) fields. This included cases where award dates preceded dispatch dates, which required corrections during the data cleaning phase. A manual reassignment of tenders to calendar years was performed to ensure temporal accuracy. However, as this was primarily based on the oldest available date for each tender, and given that TED data is not verified, there remains the possibility of minor misclassifications in identifying the exact year a CAE began using environmental criteria. That said, this influence is expected to be limited and unlikely to affect the analysis or the overall conclusions.

The first statistical difference in proportion with the *Pre* period can be noticed in 2016. The very strong p-value of $1.613e^{-4}$ confirms that the difference is significant, in addition it can be stated with 95% certainty that the percentage in 2016 is higher with a varying value of 1.75% to 5.97%. On an annual basis, the first significant jump can be seen between 2015 and 2016 with a p-value of 0.0117.

Although a strong increase in the use of environmental criteria is only visible after the transposition deadline, the announcement date (26/02/2014) remains a more suitable cutoff due to the parallel trends assumption (Section 4.4). This date marks the official policy introduction that has influenced the trends, even if the visible impact is only noticeable later. Policies often need time to be effectively implemented. The increase in green contracts after the transposition deadline is assumed to be attributed to the delay in the application of the directive. Therefore, the announcement date remains a relevant choice as a cutoff to evaluate the impact of the policy, as it provides a clear reference point.

The staggered adoption is an implication that can be addressed in future research by possibly applying the staggered DiD framework of Callaway and Sant'Anna (2021) [5], which is able to capture similar delays in adoption well.

Within the constraints and assumptions of this research, it can be claimed that the utilization of environmental award criteria in construction public procurement contracts has increased significantly since the introduction of the 2014 Public Procurement Directive. Statistical analysis has shown that, while the policy did not have an immediate effect, it significantly prompted CAEs to reconsider and substitute their conservative tendering approaches with greener alternatives.

5.2 Effect of Environmental Award Criteria on Bidding Competitiveness

The second analysis focused on estimating the impact of the treatment on bidding competitiveness, which is approximated by the number of tenders received. For each configuration (Only Green, No Reversal, Only Once), two models were used to estimate the treatment effect: A simple *Baseline* model and an *Adjusted* model. The *Adjusted* model includes additional covariates to attempt to reduce distortions from macro-economic effects, in order to increase overall robustness. More information on the used covariates is given in Section 4.4. The R-code for both models is:

```
library(fixest)
# Baseline Model
feols(
```

[13] This duration is tender and procedure specific.

```
  Y ~ post + treat + post:treat | CAE_ID, data = data_config_x)%>%
summary(cluster = ~CAE_ID)

# Adjusted Model
feols(
  Y ~ post * treat + SDG + GOVexp + covid + CPPI_growth | CAE_ID,
  data = data_config_x)%>%
summary(cluster = ~CAE_ID)
```

The one-way fixed effects model outputs are the estimated values of each provided (not excluded) coefficient. Along with their standard error clustered at the CAE_ID level, t-statistic, and p-value. These values indicate the size, precision, and statistical significance of each variable's effect. Alongside the outputs, three output metrics are printed: *RMSE*, *Adjusted R²*, and *Within R²*.

- RMSE (Root Mean Squared Error) is the average magnitude of prediction errors, measuring the distance between predicted and actual values. The RMSE gives an indication of how erroneous the predictions are. The lower this metric, the better the model fit.
- The Adjusted R is equivalent to the percentage of total variance explained by the model, adjusted for the amount of predictors. This metric punishes adding irrelevant variables, in other words punishing complexity.
- The Within R² indicates the proportion of within-unit (within-CAE in this study) variation in the outcome variable that is explained by the model.

Given the studies context and approach (panel data with fixed effects), *Within R²* is preferred over *Adjusted R²*, as the models attempt to explain changes within units, not differences between units. It is important to note that a higher R² value, does not necessarily mean the model is better. While both statistics provide an indication of how well the model fits the data, in one-way fixed effects DiD models, improving these values is considered a nice-to-have rather than a necessity. The principal purpose lies with the identification of credible treatment effects, not in maximizing predictive power.

An important remark has to be made before interpreting the regression coefficients, in connection with the log-transformation the dependent variables. In this study, the outcome variable (number of tenders received) is log-transformed but the regressors are not, necessitating a different interpretation method. Normally, the coefficients are interpreted as: a 1-unit change in regressor X leads to a β-unit change in Y. But with logged outcome variables the coefficients need to be interpreted as the percentage change in Y for a 1-unit change in regressor X [66]. For interaction terms such as $post : treat$, the estimated coefficient reflects the average percentage change in the outcome variable when a CAE transitions from untreated pre-policy state to the treated post-policy state.

To extract the exact percentage values, the following equation (found in [66]) can be applied:

$$\text{Percentage change in } Y = \left(e^{\beta_j} - 1\right) \times 100 \tag{5}$$

All Received Tenders model outputs and metrics across the different configurations are presented in Table 11 in Appendix 7.9.

Within Table 11 the focus primarily lies on the *post1:treat1* coefficient and the output metrics. *Post1:treat1* captures the DiD effect also known as the ATT (Average Treatment Effect on the Treated). Within this setting, the estimated treatment effect indicates how the change in the average number of tenders per call, before and after the adoption of environmental criteria, differs between contracting authorities that adopted green criteria and those that did not.

Across all model specifications the integration of environmental criteria is associated with statistically significant (at the 5% level) decreases in the number of tenders received per call. Observed effects suggest reductions ranging from 15.16% (= -0.1644) to as much as 23.73% (= -0.2709) depending on treatment definition and applied model. The corresponding 95% confidence intervals for these effects are [-.293 : -.036] and [-.507 : -.035] respectively.

Comparisons along the three configurations logically reveal that stricter treatment definition, which corresponds to more consistent adoption of environmental criteria, results in larger reductions. Only Green has the strongest negative effects (-20.75% : -23.73%), occasional adoption of criteria has the weakest, yet still significant impact (-15.16% : -16.27%). No Reversal, the middle-of-the-road configuration, has values between the more extreme ones (-19.37% : -20.73%).

Notably, the magnitude and significance of the *post1:treat1* interaction term remains relatively consistent within each configuration. As expected Adjusted R^2 (between 28% and 31%) and Within R^2 (between 1.5% and 2.5%) improve by including covariates, although only marginally. Therefore, it can be assumed that the treatment effect is not driven by the included covariates. In other words, the statistically significant (at 5% level) negative treatment effect is robust to controlling for macro-economic trends, thus the decrease in received tenders can be linked directly to the adoption environmental award criteria.

RMSE values between 0.47 and 0.48 are not desirable, but are acceptable predictive errors given the noisiness of the real-world EU tender data. While the model does not explain a large proportion of all variation, it credibly identifies the treatment effect, which is the main research goal.

Residuals versus fitted values plots for all model configurations are shown in Appendix 7.10. The residuals are generally symmetrically distributed around zero without strong systematic patterns, suggesting that the linear specification is appropriate.

However, a slight widening also known as a funnel shape is visible in all plots, indicating modest heteroskedasticity, with residual variance increasing at higher fitted values. Given the real-world context of noisy tendering data, the observed heteroskedasticity is not unexpected. However, it is unlikely to severely bias coefficient estimates or inference, as cluster-robust standard errors at the CAE level were applied to correct for both heteroskedasticity and intra-group correlation. The observed straight residual lines stem from the discrete nature of the outcome variable.

Within the constraints and assumptions of this research, it can be claimed that the integration of environmental award criteria in construction public procurement contracts is associated with a statistically significant (at the 5% level) decrease in the number of tenders received within treated contracting authorities. Suggesting, that green award criteria may unwantedly act as a barrier to participation within construction public procurement contracts.

Based on the one-way fixed effects DiD analysis with standard errors clustered at the CAE level, EU contracting authorities and entities that adopted green criteria after the introduction of the 2014 PP directive suffered an average decrease of approximately 15.1% to as much as 23.7% in the number of tenders received per call compared to their pre-treatment period, with estimated magnitudes depending on the configuration. With the analysis relying on within-CAE variation over time, the findings reflect changes relative to each CAE's pre-treatment behavior, rather than differences between different authorities.

Comparing configurations showed that stricter and more consistent application of environmental criteria corresponded to a greater reduction. Comparing models (baseline vs. adjusted) underlined the robustness of the results, as they persisted after adding covariates.

Within the scope of this study, adopting green criteria appears to reduce bidding competitiveness, as proxied by the number of tenders received per call. In Section 6, the implications of this finding will be discussed further. However, given the within-unit identification of treatment effects, generalization beyond the sampled CAEs must be done with caution.

5.3 Effect of Environmental Award Criteria on Project Cost

The third analysis is centered around estimating the impact of including green criteria in construction PP contracts on project costs. As the actual project cost can not be reliably found, it is approximated by the tentative contract values captured in the TED dataset.

The examination setup is identical to that discussed in Section 5.2, except for the use of a different outcome variable. However, this variable is also log-transformed, so the estimated $post1 : treat1$ interaction term coefficient, again needs to be interpreted as the average percentage change in the outcome variable (i.e. contract value) when a CAE transitions from untreated pre-policy state to the treated post-policy state.

All Contract Value model outputs and metrics across the different configurations are displayed in Table 12 in Appendix 7.11.

At first glance the presented model results indicate, that the integration of environmental award criteria correlates with a slight to moderate increase in contract values awarded. However, significance and gravity of this effect vary depending on the strictness of treatment definition and model specification. Unlike the estimated effects on the number of received tenders, there is no clear cut answer applicable to all configurations and models.

Yet, similar patterns exist across configurations, for example all baseline models show that treatment has a positive (and in the case of Only Green; statistically significant at the 5% level) impact on contract value. Additionally, the adjusted variants consistently reports slightly lower and less statistically significant estimated treatment effects compared to their respective baselines.

Another noticeable trend across all configurations relate to the reported output metrics. Across-the-board RMSE values remain fairly consistent, ranging between 0.3137 and 0.3204, suggesting comparable predictive precision. Marginally lower RMSE in adjusted models hints towards a somewhat better fit when including covariates, but the differences can be nullified. Including covariates also consistently enhances the proportion of variance explained within units (within R^2), implying these covariates capture some additional variation within contracting authorities over time. However, adjusted R^2 slightly decreases, reflecting the discussed penalty for increasing model complexity with regressors that provide limited incremental explanatory power. In general the low within R^2 values imply that most of the variation in contract values remains unexplained by treatment and controls alone.

Comparing model quality metrics of both outcome variables shows that the included covariates contribute more to explaining variation in the number of received tenders than in contract values. In the Received Tenders models, covariates slightly improve or maintain both within R^2 and adjusted R^2, while in the contract value models, they lead to a small drop in adjusted R^2 and only marginal gains in within R^2. This indicates that covariates are more informative for changes in bidding competitiveness than for changes in awarded contract values.

The Only Green baseline configuration is the only model reporting statistically significant estimated treatment effects at the 5% level. The results indicate that, with 95% confidence, transitioning from the untreated pre-policy state to the treated post-policy state is associated with an approximate 17.5% (= 0.1620 // 95% CI: [.0047:.3193]) increase in awarded contract value compared to the CAE's own pre-treatment period. By adding covariates to control for macro-economic effects, the estimated effects reduce to a positive increase of 14.6% (= 0.1362 // 95% CI: [-.0259:.2982]), which is significant at the 10% level.

The No Reversal and Only Once models report no statistically significant results. While still positive, the estimated effects are considerably lower, contract values of CAEs changing towards eco-friendlier procurement strategies are approximated to increase between 5.2% and 8.2% compared to their own pre-treatment period.

Combining these results, suggest that allowing for minor reversals in green criteria usage within CAEs substantially weakens the impact on contract valuations. Although evidence of a modest positive association between the adoption of environmental criteria and awarded contract values at the CAE level is found, this effect is sensitive to the persistence of treatment.

Once more, residuals versus fitted values plots were created for all contract value model configurations and are shown in Appendix 7.12. Across all specifications, the residuals appear symmetrically distributed around zero, with no strong evidence of nonlinearity, suggesting that the linear model form is appropriate.

Nonetheless, a mild widening (i.e. funnel shape) can be observed in several plots, particularly at higher fitted values, indicating slight heteroskedasticity where residual variance increases with predicted contract values. While some heteroskedasticity is visually apparent, its magnitude is moderate, and robust inference is ensured through the use of cluster-robust standard errors.

Based on the one-way fixed effects DiD analysis with standard errors clustered at the CAE level, EU procuring authorities and entities that adopted environmental award criteria after the introduction of the 2014 PP directive experienced an estimated increase in awarded contract values of approximately 5.2% to as much as 17.5% compared to their pre-treatment period, depending on the configuration. However, it should be noted that only the "Only Green" baseline model reports a statistically significant effect at the 5% level. As the analysis relies on within-CAE variation over time, the findings reflect changes relative to each CAE's own pre-treatment behavior, rather than differences across different authorities.

Within the scope of this study, the integration of environmental award criteria into construction procurement contracts appears to be associated with a modest increase in awarded contract values, particularly when the criteria are applied consistently over time. However, given the limited statistical significance across configurations, these findings should be interpreted cautiously. In Section 6, the broader inferences of these findings will be further explored. As with the bidding competitiveness results, generalization beyond the observed CAEs must be done with care due to the usage of within-CAE variation for causal identification.

In analyzing both awarded contract values and bidding competitiveness, this study differentiated contracting authorities through various treatment configurations based on their consistency of green criteria usage. While this approach captures differences in adoption patterns, the current framework does not explicitly model the differential effect between fully pure green users and mixed adopters within a single specification. This point will be revisited in Section 6, where the scope and limitations of the study are further elaborated.

6 Discussion

The relevance of this thesis lies in critically researching and reviewing the impact of the implementation of the 2014 directive of the European Parliament and the Council on public procurement, more specifically focused on the impact of GPP within the construction sector.

GPP is presented as a strategic instrument to align public spending with environmental objectives [44][49] to create a breeding ground for sustainable practices and innovation within the safety of the Single Market with its core principles of non-discrimination and free competition [51]. The EC aims to balance economic growth with environmental responsibility across the EU by weaving sustainability into these principles [42][49]. However, relevant literature hypothesizes that GPP inadvertently hampers these principles by introducing entry barriers that may favor certain suppliers more than others [67]. As such, it unintentionally reduces supplier diversity and compromises the integrity of the internal market by violating its principles, ensuring equal access, transparency, and competition within public procurement. Hence, the question arises whether the dual purpose of promoting sustainability and preserving competition and non-discrimination can be achieved in practice.

Existing literature on the limitations of GPP as a strategic policy for achieving environmental objectives, discussed extensively in Sections 3.4 and 3.5, highlights the need for a more pragmatic study assessing the impact of implementing green procurement practices. Therefore, this thesis empirically investigates the integration of environmental award criteria, stimulated by the 2014 PP directive, in construction PP within the EU and how these criteria impact procurement outcomes, specifically, bidding competitiveness and project cost.

The findings of the thesis contribute to the ongoing academic and policy debate by providing empirical evidence on the extent of misalignment between the observable impacts of utilizing green award criteria and the fundamental principles of the EU internal market.

The detailed results, regarding the impact on bidding competitiveness, are discussed in Section 5.2. In summary, within the analyzed sample of CAEs placing construction tenders, those that adopted green award criteria after the announcement of the 2014 PP directive experienced an average decrease between approximately 15.1% and 23.7% in the number of tenders received per call compared to their pre-policy period. This effect was statistically significant at the 5% level and remained robust after controlling for macroeconomic variables across all treatment configurations.

These impacts confirm existing concerns and hypotheses formulated by Drake et al. (2024) [14], Lundberg et al. (2015) [45], and Van Assche et al. (2024) [67] that voluntary and heterogeneous integration of environmental award criteria would impose unwanted tender participation barriers. This spillover effect implies that the core principles of the Single Market are violated, putting tension on the belief that GPP is a valid environmental policy tool.

A significant negative causal relationship is observed between the application of environmental award criteria and the number of tenders received. However, upon critically reflecting on the estimates and considering the legal definition of award criteria, it is noteworthy that such criteria do not explicitly or legally prohibit suppliers who cannot meet them from placing a tender offer. Consequently, other mechanisms might influence suppliers' decisions to enter the bidding process. Withdrawal is likely economically or strategically motivated. Several mechanisms that might indirectly discourage participation include:

- Higher (perceived) compliance costs: Suppliers who cannot easily meet green criteria might expect lower evaluation scores and a lower chance of winning. Thus, opting for the rational choice not to waste resources and time preparing a formal bid.
- Strategic self-selection: Suppliers aware of their competitive disadvantage in green criteria or those that perceive themselves as non-green may voluntarily opt out of competing.
- Asymmetric information: Suppliers might fear hidden barriers or perceive (even if incorrectly) that the procedure favors certain green-qualified firms.

As a result, even without formal exclusion, environmental award criteria function as a soft barrier to market entry, reducing bidding competitiveness. The possible presented causes remain purely speculative within the scope of this study and require further examination. Exploring which potential underlying considerations influence supplier participation into GPP contracts presents a valuable direction for future research.

The results show a more nuanced picture of the impact on awarded contract values. The regression results partially substantiate the statement made by Carreras (2023) [6] that the effect of GPP on contract prices is difficult to generalize due to the contextual heterogeneity of public procurement tenders.

Estimates show that CAEs that adopted green award criteria after the introduction of the 2014 PP directive experienced modest increases in project costs compared to their pre-treatment period. Increases ranging from 5.2% to 17.5% were observed, albeit that statistical significance (at 5% in baseline and 10% in adjusted) was only achieved in the most stringent treatment configuration. This disproves part of the statement by Keaveny and Butler (2014) [41] that GPP has a misconception of higher associated costs. The results prove that, in specific settings, it significantly drives costs upwards. However, more research is needed to substantiate this claim more concretely at the individual tender level.

The thesis's findings on project costs call into question the demand Drake et al. (2024) [14] made for mandatory utilization of green criteria, as it could be associated with price premiums beyond acceptable levels. This thesis defines these acceptable levels as the point where the price premium of greener tenders outweighs their environmental benefits, or where the gap between green and lowest-price contracts becomes unreasonably wide. In line with Drake et al. (2024), such levels would violate the condition that the public sector must be willing to pay a premium for less polluting options, and that at least one "brown" supplier would find it worthwhile to invest in compliance with the environmental criterion. The empirical results of this thesis create a foundation to solicit further investigation that can evaluate the willingness and capacity of the public sector to absorb these increases. This discussion becomes increasingly relevant as the upcoming Ecodesign for Sustainable Products Regulation (ESPR) introduces mandatory sustainability criteria for manufacturers, and enables mandatory GPP rules for specific product categories [23].

Regarding the link between both outcome variables, it is known from traditional lowest-price tendering that fewer offers typically lead to higher awarded contract values, as increased competition tends to drive prices downward [14][45]. However, no clear relationship can be established based on the estimated effects. It remains unclear whether bidding competitiveness and contract values are still systematically linked in sustainable procurement under the guidelines of the 2014 PP directive. In particular, due to using the Most Economically Advantageous Tender (MEAT) principle, contracting authorities are not obligated to select the lowest bid. Consequently, even when many offers are received, authorities may still opt for higher-priced tenders that better meet qualitative and sustainability criteria.

To fully assess whether the relationship between competition and prices persists when environmental award criteria are applied, future research would need to explicitly model the interaction between the number of offers and awarded contract values. Given the low data quality, this was deemed not achievable within this thesis. Future research would require richer, more standardized data, capturing detailed evaluation scores and award decisions to separate price effects from changes in competition after applying sustainability considerations during supplier selection.

As mentioned in Section 5, this study differentiated CAEs between three intersecting treatment groups, through various treatment configurations based on their consistency of green criteria usage. In doing so, the fixed effect DiD model captures differences in adoption patterns. However, the current framework does not explicitly model the differential effect between fully pure green users and mixed adopters within a single specification.

Introducing an interaction term between treatment, post-treatment, and a *pure green* indicator ($post : treat : pure$) could help future research more precisely estimate whether fully consistent adoption of environmental criteria leads to distinct outcome impacts compared to partial adoption strategies. Although an exploratory model incorporating this interaction suggests no statistically significant difference between pure and mixed adopters, a more comprehensive evaluation of this aspect is recommended in future work to validate these findings and to assess potential heterogeneity in treatment effects across different adoption intensities.

Despite the intriguing findings from this study, some shortcomings affect the results' rigor and interpretation. These can be addressed in future research. Multiple methodological challenges shaped the study's design choices. First, due to data quality issues within the TED datasets, analysis at the individual tender lot level was not feasible. This inability constitutes a significant limitation, as tenders almost always cover multiple lots, each with varying size and complexity. By aggregating at the call for tender level and, as such, making harmonization assumptions to eliminate discrepancies, such as the complete call for tender being classified as green if at least one lot or sub-tender is green, the heterogeneity is reduced. Still, essential aspects possibly influencing bidding competitiveness and project costs are obscured, weakening the power and accuracy of any causal claims.

Secondly, TED data is known for its structural complexity, missing values, redundancy, and inconsistencies [2][55][59][61]. Crucial information fields for this thesis, such as contract values and the number of offers, are often incomplete or entered with placeholder values, an inherent limitation in the reliability of this procurement analysis, and therefore deemed a primary constraint to analytical capability by Ackermann et al. (2019) [2]

Another limitation lies in the covariates used. Ideally, time-variant explanatory variables would be defined uniquely at the CAE level. However, given the scarcity of such variables, covariates were sourced at the country level instead. Although these macroeconomic indicators add some explanatory value, a comparison of model output metrics shows they contribute more meaningfully to explaining variation in bidding competitiveness than in awarded contract values. Thus, external factors influence supplier participation more strongly than pricing behavior. Future research at the CAE level could manually search for, identify, and include characteristics by using the provided identifying information. For the scope of this thesis, this was considered too time-intensive.

Additionally, important factors, such as the expected duration and size of the contracts, were not provided. These likely influence both outcome variables heavily and would aid in extracting the individual effect of applying environmental criteria on the outcomes of the procuring process. Yet, in this study's current two-period CAE level configuration, these covariates would be less helpful, as the tender level heterogeneity is absorbed, possibly creating false correlations. Another related methodological constraint is the inability to account for variation in procurement procedure types. Procedures may be associated with different levels of supplier participation, potentially influencing the number of bids received. However, since the analysis aggregates data at the CAE level across multiple tenders each potentially applying a different procedure it was not feasible to include this variable in the model due to the heterogeneity of procedure types across observations. As a result, no claims can be made within this thesis about the relationship between the number of tenders received and the specific procurement procedure applied.

The two-period CAE-level setup, adopted for this analysis, presented strengths and weaknesses. Solely using a binary post-treatment period indicator, the model does not absorb time-fixed effects in the classical two-way fixed effects (TWFE) sense. As a result, time-specific (macroeconomic) shocks that uniformly affect all contracting authorities are less effectively captured. Furthermore, because the analysis coalesces observations across many periods, the model may be less robust to unobserved time-varying factors influencing procurement dynamics. The used framework captured general pre- and post-treatment dynamics while avoiding complications associated with staggered adoption and irregular observation timing. An attempt to capture these time-variant effects was made using covariates such as the COVID indicator, reflecting the relative intensity of pandemic-related procurement activity, and the CPPI_growth variable, capturing the one-year-delayed growth of the construction sector. However, results revealed that these variables exhibited limited statistical power.

Several design assumptions were made to sharpen the analytical focus to achieve the desired scope and make the analysis operationally feasible. Specifically, it was assumed that the 2014 PP Directive's primary environmental intention justified restricting the sample to calls that used or did not use green award criteria. Non-green calls for tenders applying the MEAT principle to award contracts were excluded to zero in on the effect of environmental criteria on procurement outcomes. Including these non-green contracts could lead to an under- or overestimation of the impact. The decision to exclude these non-green contracts is therefore justified. However, it also introduces a residual risk: CAEs that previously utilized non-green MEAT criteria might still display behavioral changes in their subsequent lowest-price contracts, a phenomenon not fully captured in this study.

Another shortcoming relates to the identification of green contracts. Testa et al. (2016) [63] mention that environmental criteria can be included across multiple components. Given the scale and limited availability of sufficient identifiers of environmental practices within the dataset, only the award criteria field was used to identify green contracts in this study. Therefore, green practices implemented elsewhere may have been overlooked. As such, the green classification might underestimate the prevalence of sustainable procurement practices, representing another potential source of measurement error.

From a statistical perspective, although fixed-effects Difference-in-Differences (DiD) estimation provides strong internal validity by controlling for unobserved time-invariant heterogeneity [31][47], the generalizability of results beyond the sampled CAEs remains limited. While mild heteroskedasticity, particularly at higher predicted values, was visible in residual plots, its effect on inference is mitigated through cluster-robust standard errors at the CAE level. However, some heteroskedasticity may remain and cannot be entirely ruled out.

Nevertheless, given the practical constraints, the research design adopted here, focusing on CAE-level aggregation, fixed effects, and a two-period Difference-in-Differences approach, offers a sound and defensible estimation strategy for identifying the causal impact of environmental award criteria on procurement outcomes in the EU construction sector.

This thesis provides empirical evidence on the implications of including environmental award criteria within the context of EU construction procurement. Using a two-period one-way fixed effects Difference-in-Differences model with CAE-level aggregation, the analysis revealed that CAEs that switched to green procurement strategies experienced a statistically significant (at the 5% level) decrease in offers, ranging from 15.1% to upwards of 23.7%. The decrease in the number of tenders received reflects a reduction in bidding competitiveness. This decline cannot be directly attributed to the pure inclusion of environmental criteria, as these criteria do not explicitly or legally prohibit suppliers from submitting bids. Most likely, the drop in participation is influenced by other factors that have not yet been identified. This thesis can serve as a raision d'être for further research on these still unknown causes affecting supplier participation in GPP contracts. The findings also disclose a more subtle impact on awarded contract values. While the regression results suggest that green procurement leads to modest increases in project costs (between 5% and 17.5%), statistical significance was only achieved in the most stringent treatment configuration. These findings suggest that GPP can, in specific settings, significantly drive up costs, though the results are not definitive, and further research is needed to substantiate this claim more concretely at the tender level. The results can not identify a clear relationship between competitiveness and project costs in the context of GPP. The use of the MEAT principle complicates the direct link between the number of tenders received and the awarded contract values, as CAEs are not required to select the lowest bid.

While the integration of green criteria is intended to advance sustainability within the boundaries and safety of the Single Market, the findings suggest that it unintentionally introduces barriers to competition, disrupting the very principles that underpin the EU's internal market. Therefore, the thesis challenges the EU's belief and understanding of the PP Directive's impact and potential, showing that the intended balance between promoting sustainability and ensuring fair economic growth has not been achieved. The integration of environmental criteria, while aligning with environmental goals, inadvertently significantly hampers competition and might increase project costs beyond acceptable levels. Based on the provided insights, this thesis advises that the practical application of the Directive needs to be reviewed, calling for a reevaluation of its implementation to better align sustainability objectives with the principles of free market competition. Future research could explore instances where reduced competition in GPP is an intentional trade-off made to achieve broader environmental objectives. Understanding these connections would help clarify whether the observed decline in bidding competitiveness is an unintended consequence or a justifiable policy choice. However, identifying such cases and their motivations is beyond the scope and feasibility of this thesis, which takes a data-driven and outcome-focused approach.

References

1. Acconia. Sustainable building materials. https://www.activesustainability.com/construction-and-urban-development/sustainable-building-materials/.
2. Roderick Ackermann, Margarita Sanz, and Antonio Sanz. Gaps and errors in the TED database. February 2019.
3. Mohammed Zajeer Ahmed, Cathal O'Donoghue, and Patrick McGetrick. Green public procurement in construction: A systematic review. *Cleaner and Responsible Consumption*, 15:100234, December 2024.
4. Omar Binshakir, Lulwa AlGhanim, Afreen Fathaq, Alsharif Mohamed AlHarith, Salma Ahmed, and Sameh El-Sayegh. Factors Affecting the Bidding Decision in Sustainable Construction. *Sustainability*, 15(19):14225, 2023.
5. Brantly Callaway and Pedro H. C. Sant'Anna. Difference-in-Differences with multiple time periods. *Journal of Econometrics*, 225(2):200–230, December 2021.
6. Enrique Carreras. Sustainable Public Procurement, Bidding Behavior and Costs: Evidence from Spanish Contracts. 2023.
7. Columbia Public Health. Difference-in-Difference Estimation, August 2016. https://www.publichealth.columbia.edu/research/population-health-methods/difference-difference-estimation/.
8. Garcia Jorge Cristobal, Dario Caro, Gillian Foster, Giulia Pristera, Federico Gallo, and Davide Tonini. Techno-economic and environmental assessment of construction and demolition waste management in the European Union, 2024. https://publications.jrc.ec.europa.eu/repository/handle/JRC135470.
9. Clément de Chaisemartin and Xavier D'Haultfœuille. Two-Way Fixed Effects Estimators with Heterogeneous Treatment Effects. *The American Economic Review*, 110(9):2964–2996, 2020.
10. Directorate-General for Climate Action. Let's get our heads out of the sand – it's time we got serious about preparing for the climate crisis - European Commission. https://climate.ec.europa.eu/news-your-voice/news/lets-get-our-heads-out-sand-its-time-we-got-serious-about-preparing-climate-crisis-2024-12-12_en.
11. Nicholas Dodd, Elena Garbarino, and Miguel Gama Caldas. Green Public Procurement Criteria for Office Building Design, Construction and Management, 2016.
12. Shane Donatello, Aleksandra Arcipowska, Zahara Perez, and Angela Ranea. EU Green Public Procurement (GPP) criteria for the design, construction, renovation, demolition and management of buildings. February 2022.
13. Dacian Dragos and Bogdana Neamtu. Sustainable Public Procurement: Life-Cycle Costing in the New EU Directive Proposal. *European Procurement & Public Private Partnership Law Review*, 8(1):19–30, 2013.

14. Samielle Drake, Johan Lundberg, and Sofia Lundberg. Market response to environmental policy via public procurement: An empirical analysis of bids and prices. *Journal of Environmental Management*, 365, June 2024.
15. Peter Džupka and Matúš Kubák. Sustainable Public Procurement in Central European Countries. Can It Also Bring Savings? *Sustainability*, 12(21):9241, 2020.
16. EU CAP NETWORK. Learning portal - Propensity score matching-difference in differences (PSM-DiD) method | EU CAP Network. https://eu-cap-network.ec.europa.eu/training/evaluation-learning-portal/learning-portal-propensity-score-matching-difference-differences-psm-did-method_en.
17. European Commission. Access to public procurement | Single Market Scoreboard. https://single-market-scoreboard.ec.europa.eu/business-framework-conditions/public-procurement_en.
18. European Commission. Buildings and construction. https://single-market-economy.ec.europa.eu/industry/sustainability/buildings-and-construction_en.
19. European Commission. Climate action and the Green Deal. https://commission.europa.eu/strategy-and-policy/priorities-2019-2024/european-green-deal/climate-action-and-green-deal_en.
20. European Commission. Construction Products Regulation (CPR). https://single-market-economy.ec.europa.eu/sectors/construction/construction-products-regulation-cpr_en.
21. European Commission. Contract Procedures for EU External Action: A practical guide. https://wikis.ec.europa.eu/spaces/ExactExternalWiki/pages/152798726/5.+Works+Contracts.
22. European Commission. Delivering the European Green Deal. https://commission.europa.eu/strategy-and-policy/priorities-2019-2024/european-green-deal/delivering-european-green-deal_en.
23. European Commission. Ecodesign for Sustainable Products Regulation. https://commission.europa.eu/energy-climate-change-environment/standards-tools-and-labels/products-labelling-rules-and-requirements/ecodesign-sustainable-products-regulation_en#:~:text=Overview,can%20significantly%20impact%20the%20environment.
24. European Commission. Energy Performance of Buildings Directive. https://energy.ec.europa.eu/topics/energy-efficiency/energy-efficient-buildings/energy-performance-buildings-directive_en.
25. European Commission. The European Green Deal. https://commission.europa.eu/strategy-and-policy/priorities-2019-2024/european-green-deal_en.
26. European Commission. Green Public Procurement. https://green-business.ec.europa.eu/green-public-procurement_en.
27. European Commission. Public tendering rules. https://europa.eu/youreurope/business/selling-in-eu/public-contracts/public-tendering-rules/index_en.htm.
28. European Union. Competition - EUR-Lex. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=legissum:competition.
29. European Union. Directive 2004/18 - Coordination of procedures for the award of public works contracts, public supply contracts and public service contracts - EU monitor. https://eur-lex.europa.eu/eli/dir/2004/18/oj/eng.
30. European Union. Principle of proportionality - EUR-Lex. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=legissum:proportionality.
31. Christophe Frey, Christoph Scheuch, Stefan Voigt, and Patrick Weiss. Fixed Effects and Clustered Standard Errors with R. https://www.tidy-finance.org/r/fixed-effects-and-clustered-standard-errors.html.
32. Jose Luis Fuentes-Bargues, Pablo Sebastian Ferrer-Gisbert, Mª Carmen González-Cruz, and María Jose Bastante-Ceca. Green Public Procurement at a Regional Level. Case Study: The Valencia Region of Spain. *International Journal of Environmental Research and Public Health*, 16(16):2936, 2019.
33. Jose Luis Fuentes-Bargues, Mª Carmen Gonzalez-Cruz, and Cristina Gonzalez-Gaya. Environmental Criteria in the Spanish Public Works Procurement Process. *International Journal of Environmental Research and Public Health*, 14(2):204, February 2017.
34. Manuel J. García Rodríguez, Vicente Rodríguez Montequín, Francisco Ortega Fernández, and Joaquín M. Villanueva Balsera. Public Procurement Announcements in Spain: Regulations, Data Analysis, and Award Price Estimator Using Machine Learning. *Complexity*, 2019, 2019.

35. Melissa M. Garrido, Amy S. Kelley, Julia Paris, Katherine Roza, Diane E. Meier, R. Sean Morrison, and Melissa D. Aldridge. Methods for Constructing and Assessing Propensity Scores. *Health Services Research*, 49(5):1701–1720, 2014.
36. Jolien Grandia and Dylan Voncken. Sustainable Public Procurement: The Impact of Ability, Motivation, and Opportunity on the Implementation of Different Types of Sustainable Public Procurement. *Sustainability*, 11(19):5215, 2019.
37. Fatima Hafsa, Nicole Darnall, and Stuart Bretschneider. Estimating the True Size of Public Procurement to Assess Sustainability Impact. *Sustainability*, 13(3):1448, 2021.
38. Qi Han, Sytske Blaauwbroek, and Pei-Hsuan Lee. Environment Performance Criteria as Means for Green Public Procurement. In John R. Littlewood, Lakhmi Jain, and Robert J. Howlett, editors, *Sustainability in Energy and Buildings 2023*, pages 11–22. Springer Nature, Singapore, 2024.
39. Harry Haupt. Seemingly unrelated and fixed-effect panel regressions: collinearity and singular dispersion, November 2020.
40. Nick Huntington-Klein. *Chapter 18 - Difference-in-Differences | The Effect*. 2021.
41. Myles Keaveney and Paul Butler. An Analysis of the Barriers to and Drivers of Green Public Procurement in Achieving a More Sustainable Construction Industry. *Conference papers*, March 2014.
42. Sanjay Kumar. Brief Overview of Sustainable Public Procurement. In Sanjay Kumar, editor, *Understanding Sustainable Public Procurement: Reflections from India and the World*, pages 1–16. Springer International Publishing, Cham, 2022.
43. Irene Litardi, Gloria Fiorani, and Daniele Alimonti. The State of the Art of Green Public Procurement in Europe: Documental Analysis of European Practices. In Sandro Brunelli and Emiliano Di Carlo, editors, *Accountability, Ethics and Sustainability of Organizations: New Theories, Strategies and Tools for Survival and Growth*, pages 175–192. Springer International Publishing, Cham, 2020.
44. Sofia Lundberg and Per-Olov Marklund. Green public procurement and multiple environmental objectives. *Economia e Politica Industriale*, 45(1):37–53, March 2018.
45. Sofia Lundberg, Per-olov Marklund, Elon Strömbäck, and David Sundström. Using public procurement to implement environmental policy: an empirical analysis. *Environmental Economics and Policy Studies*, 17(4):487–520, October 2015.
46. Lennard Michaux and Joris Gruyters. Life Cycle Costing: The Final Step Towards a True Rule of Reason in Public Procurement Law? *European Procurement & Public Private Partnership Law Review*, 15(1):61–69, 2020.
47. Jonathan Mummolo and Erik Peterson. Improving the Interpretation of Fixed Effects Regression Results. *Political Science Research and Methods*, 6(4):829–835, October 2018.
48. Lela Mélon. More Than a Nudge? Arguments and Tools for Mandating Green Public Procurement in the EU. *Sustainability*, 12(3):988, 2020.
49. Peter Nemec, Róbert Štefko, and Matúš Kubák. Better Environmental Value Using Public Procurement: Evidence from Visegrad Group Countries. *Finance a Uver*, 72(4):296–327, 2022.
50. Astrid Nilsson Lewis, Kaidi Kaaret, Eileen Torres Morales, Evelin Piirsalu, and Katarina Axelsson. Green public procurement : a key to decarbonizing construction and road transport in the EU. *SEI report*, 2023.
51. Antti Palmujoki, Katriina Parikka-Alhola, and Ari Ekroos. Green Public Procurement: Analysis on the Use of Environmental Criteria in Contracts. *Review of European Community & International Environmental Law*, 19(2):250–262, July 2010.
52. Katriina Parikka-Alhola and Ari Nissinen. Environmental Impacts and the Most Economically Advantageous Tender in Public Procurement. *Journal of Public Procurement*, 12(1):43–80, 2012.
53. Brigitte Pircher. EU Public Procurement Policy During Covid-19: A Turning Point for Legitimate EU Governance? *Politics and Governance*, 10(3):131–142, 2022.
54. Kleoniki Pouikli. Towards mandatory Green Public Procurement (GPP) requirements under the EU Green Deal: reconsidering the role of public procurement as an environmental policy tool. *ERA Forum*, 21(4):699–721, January 2021.
55. Eric Prier, Palina Prysmakova, and Clifford P. McCue. Analysing the European Union's Tenders Electronic Daily: possibilities and pitfalls. *International Journal of Procurement Management*, 11(6):722–747, January 2018.

56. Jordi Rosell. Did governments neglect the environment during the COVID-19 pandemic? An empirical analysis of green public procurement. *Economic Analysis and Policy*, 80:835–849, December 2023.
57. Jordi Rosell. Green Public Procurement in Spain *. *Hacienda Publica Espanola*, (245):95–117, 2023.
58. Paul R. Rosenbaum and Donald B. Rubin. The Central Role of the Propensity Score in Observational Studies for Causal Effects. *Biometrika*, 70(1):41–55, 1983.
59. André Sapir, Tom Schraepen, and Simone Tagliapietra. Green Public Procurement: A Neglected Tool in the European Green Deal Toolbox? *Intereconomics*, 57(3):175–178, May 2022.
60. Timothy Simcoe and Michael W. Toffel. Government green procurement spillovers: Evidence from municipal building policies in California. *Journal of Environmental Economics and Management*, 68(3):411–434, November 2014.
61. Ahmet Soylu, Óscar Corcho, Brian Elvesæter, Carlos Badenes-Olmedo, Francisco Yedro-Martínez, Matej Kovacic, Matej Posinkovic, Mitja Medvešček, Ian Makgill, Chris Taggart, Elena Simperl, Till C. Lech, and Roman Dumitru. Data Quality Barriers for Transparency in Public Procurement. *Information*, 13(2):99, 2022.
62. Sönnich Dahl Sönnichsen and Jesper Clement. Review of green and sustainable public procurement: Towards circular public procurement. *Journal of Cleaner Production*, 245:118901, February 2020.
63. Francesco Testa, Paolo Grappio, Natalia M. Gusmerotti, Fabio Iraldo, and Marco Frey. Examining green public procurement using content analysis: existing difficulties for procurers and useful recommendations. *Environment, Development and Sustainability*, 18(1):197–219, February 2016.
64. Tünde Tátrai and Orsolya Diófási-Kovács. European Green Deal – the way to Circular Public Procurement. *ERA Forum*, 22(3):523–539, September 2021.
65. Tünde Tátrai, Gyöngyi Vörösmarty, and Péter Juhász. Intensifying Competition in Public Procurement. *Public Organization Review*, 24(1):237–257, March 2024.
66. UCLA: Statistical Consulting Group. Introduction to SAS. https://stats.oarc.ucla.edu/other/mult-pkg/faq/general/faqhow-do-i-interpret-a-regression-model-when-some-variables-are-log-transformed/.
67. Ari Van Assche, Ali Arsalan Pasha, Lucian Cernat, and Hinrich Voss. From the editor – Governments as buyers: the international business implications of public procurement. *Journal of International Business Policy*, 7(2):133–146, June 2024.
68. Qi Wang, Renjie Zhang, and Junqi Liu. Price/time/intellectual efficiency of procurement: Uncovering the related factors in Chinese public authorities. *Journal of Purchasing and Supply Management*, 26(3):100622, June 2020.
69. Chunling Yu, Toru Morotomi, and Haiping Yu. What Influences Adoption of Green Award Criteria in a Public Contract? An Empirical Analysis of 2018 European Public Procurement Contract Award Notices. *Sustainability*, 12(3):1261, 2020.

7 Appendices

7.1 Appendix 1: Consolidation Function

Listing 1.1: Consolidate pre and post values per CAE

```

get_mode <- function(x) {
  ux <- unique(x)
  tab <- tabulate(match(x, ux))
  mode_val <- ux[tab == max(tab)]
  mode_val[1]
}

data <- data %>%
  group_by(CAE_ID, cutter) %>%
  summarise(
    PRICE = mean(PRICE, na.rm = TRUE),
    NUMBER_OFFERS = mean(NUMBER_OFFERS, na.rm = TRUE),
    SDG = mean(SDG, na.rm = TRUE),
    spillover = mean(spillover, na.rm = TRUE),
    GOVexp = mean(GOVexp, na.rm = TRUE),
    covid = mean(covid, na.rm = TRUE),
    CPPI_growth = mean(CPPI_growth, na.rm = TRUE),

    region = get_mode(region),
    COUNTRY = get_mode(COUNTRY),
    CAE_TYPE = get_mode(CAE_TYPE),
    CAE_TYPE_grouped = get_mode(CAE_TYPE_grouped),

    .groups = "drop"
  )
```

7.2 Appendix 2: Interaction terms parallel trends

Table 5: Parallel Trends Assumption Check: Joint p-values and Yearly Interaction Terms for Contract Prices

Cut-off Date	Configuration	2013: treat (p)	2014: treat (p)	2015: treat (p)	2016: treat (p)	Joint F-statistic	Joint p-value
2014-01-01	Only Green	0.7533	/	/	/	0.0989	0.7533
	No Reversal	0.7890	/	/	/	0.0716	0.7891
	Only Once	0.6440	/	/	/	0.2137	0.6440
2014-02-26	Only Green	0.6550	0.1940	/	/	0.8501	0.4275
	No Reversal	0.6830	0.1510	/	/	1.0338	0.3559
	Only Once	0.5310	0.7760	/	/	0.2023	0.8164
2015-01-01	Only Green	**0.0902***	0.8988	/	/	2.0529	0.1286
	No Reversal	0.4066	0.5905	/	/	0.3471	0.7068
	Only Once	0.1938	0.7128	/	/	1.7055	0.1819
2016-01-01	Only Green	0.5340	0.4040	0.4580	/	1.0089	0.3877
	No Reversal	0.8778	0.8026	0.5680	/	0.1264	0.9445
	Only Once	0.4069	0.5974	0.8640	/	0.7495	0.5225
2016-04-18	Only Green	0.5900	0.4690	0.7390	0.9490	0.4898	0.7433
	No Reversal	0.9679	0.8159	0.3891	0.5736	0.3237	0.8622
	Only Once	0.6538	0.4619	0.8132	0.2136	0.8304	0.5056
2017-01-01	Only Green	0.4040	0.1790	0.2940	0.8870	0.6141	0.6525
	No Reversal	0.7170	0.9230	0.9370	**0.0870***	1.1248	0.3429
	Only Once	0.7040	0.5430	0.3280	0.3700	0.7668	0.5467

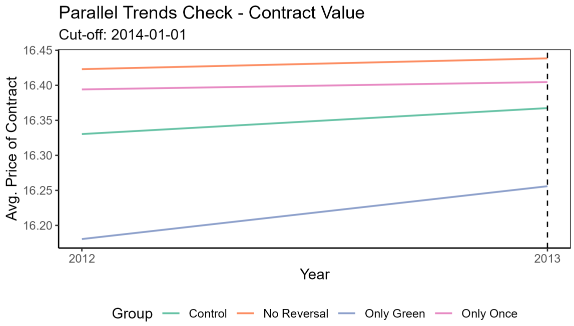
Notes: $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$

Table 6: Parallel Trends Assumption Check: Joint p-values and Yearly Interaction Terms for Tenders Received.

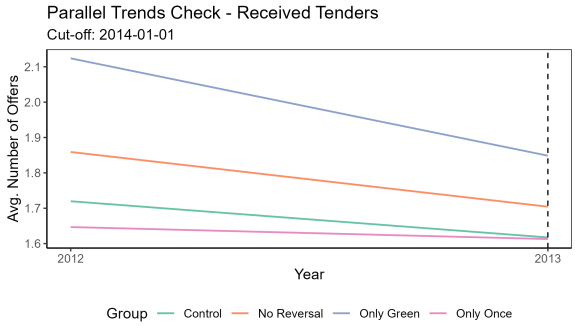
Cut-off Date	Configuration	2013: treat (p)	2014: treat (p)	2015: treat (p)	2016: treat (p)	Joint F-statistic	Joint p-value
2014-01-01	Only Green	0.3291	/	/	/	0.9532	0.3291
	No Reversal	0.6562	/	/	/	0.1982	0.6562
	Only Once	0.3942	/	/	/	0.7262	0.3942
26/02/2014	Only Green	0.2330	0.8059	/	/	0.8729	0.4179
	No Reversal	0.5757	0.4375	/	/	0.6329	0.5311
	Only Once	0.4949	**0.0377****	/	/	2.1676	0.1147
2015-01-01	Only Green	0.6732	0.2537	/	/	0.6651	0.5143
	No Reversal	0.3517	0.7793	/	/	0.4836	0.6166
	Only Once	0.5830	0.2990	/	/	0.5393	0.5832
2016-01-01	Only Green	0.6115	0.1455	0.5448	/	1.5649	0.1958
	No Reversal	0.7929	0.5990	0.4225	/	0.7928	0.4978
	Only Once	**0.0910***	0.1415	0.9974	/	1.8361	0.1338
2016-04-18	Only Green	0.8031	**0.0285****	0.2285	0.5768	2.3917	**0.0487****
	No Reversal	0.9308	0.6787	0.4193	**0.0809***	1.4135	0.2268
	Only Once	0.1870	0.2479	0.7513	**0.0100*****	2.6572	**0.0312****
2017-01-01	Only Green	0.4955	0.3677	0.9755	0.5878	0.9428	0.4380
	No Reversal	**0.0428****	0.1125	0.9914	**0.0341****	2.6642	**0.0309****
	Only Once	**0.0070*****	**0.0542***	0.9352	**0.0376****	3.4044	**0.0087*****

Notes: $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$

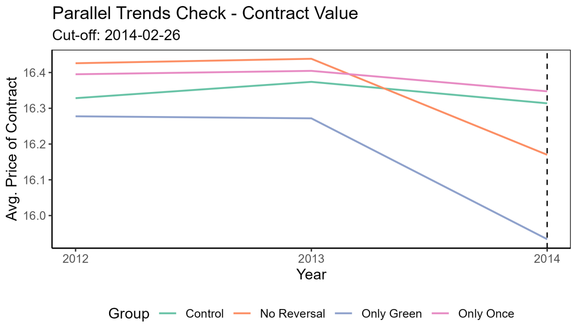
7.3 Appendix 3: Visual Plots Parallel Trends



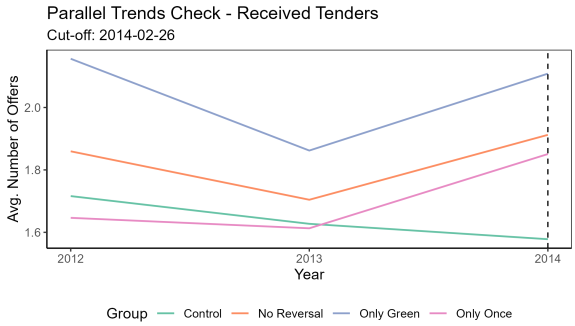
(3.1) Contract Value - Cut-off: 2014-01-01 - Trends compared across configurations



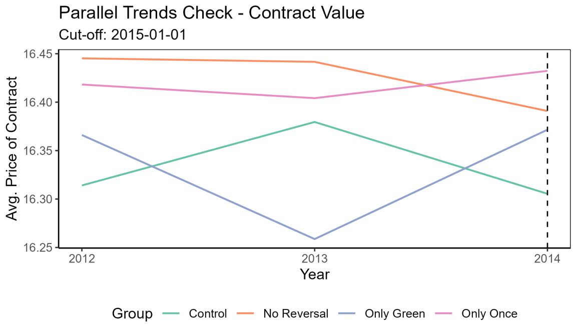
(3.2) Received Tenders - Cut-off: 2014-01-01 - Trends compared across configurations



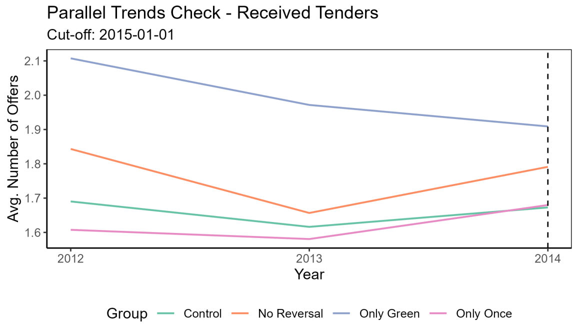
(3.3) Contract Value - Cut-off: 2014-02-26 - Trends compared across configurations



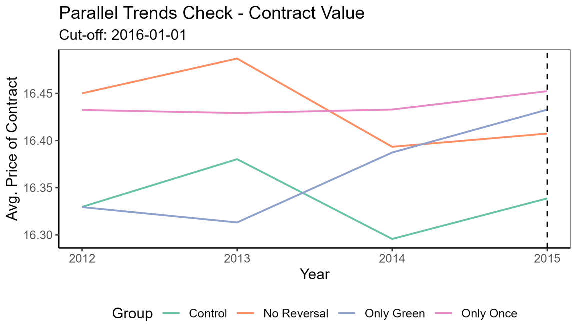
(3.4) Received Tenders - Cut-off: 2014-02-26 - Trends compared across configurations



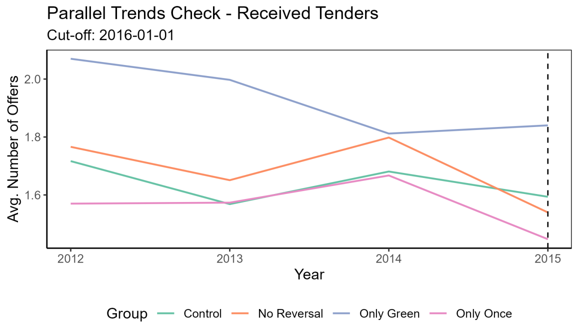
(3.5) Contract Value - Cut-off: 2015-01-01 - Trends compared across configurations



(3.6) Received Tenders - Cut-off: 2015-01-01 - Trends compared across configurations

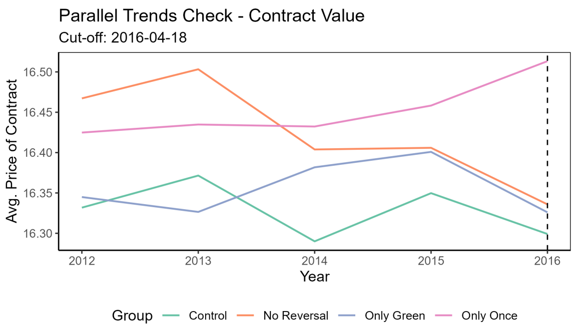


(3.7) Contract Value - Cut-off: 2016-01-01 - Trends compared across configurations

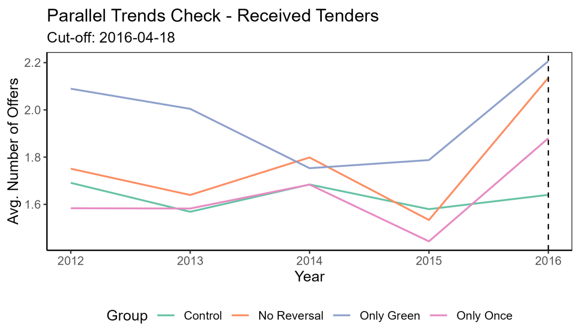


(3.8) Received Tenders - Cut-off: 2016-01-01 - Trends compared across configurations

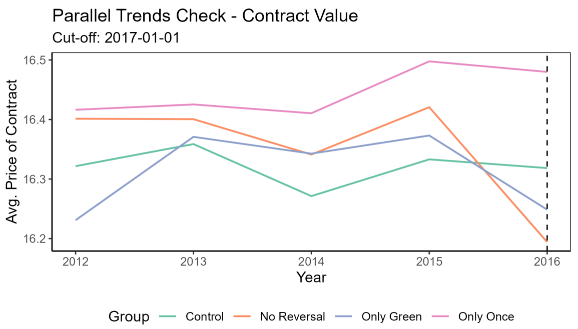
Fig. 3: Parallel trends visual plots (Part 1 of 2)



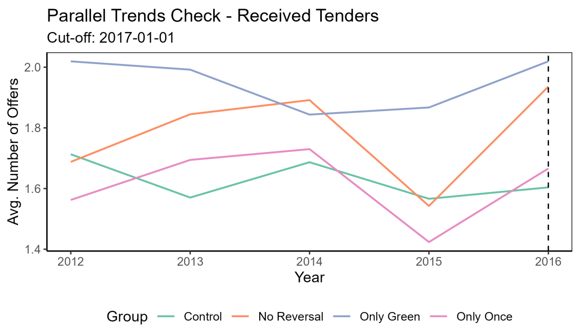
(3.9) Contract Value - Cut-off: 2016-04-18 - Trends compared across configurations



(3.10) Received Tenders - Cut-off: 2016-04-18 - Trends compared across configurations



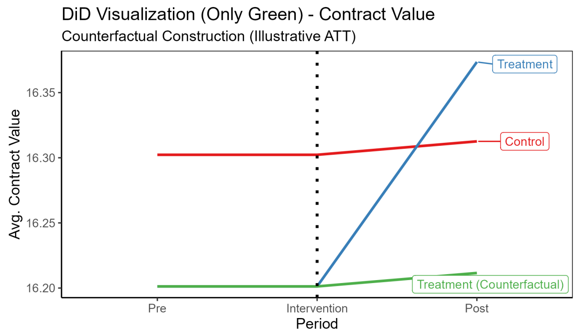
(3.11) Contract Value - Cut-off: 2017-01-01 - Trends compared across configurations



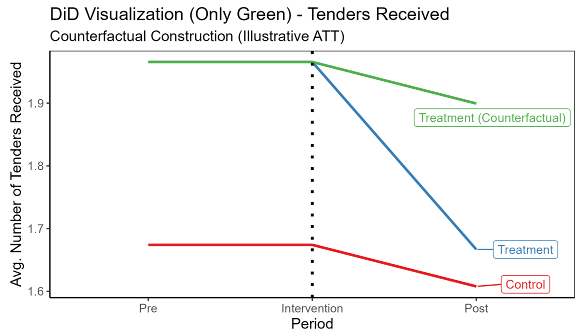
(3.12) Received Tenders - Cut-off: 2017-01-01 - Trends compared across configurations

Fig. 3: Parallel trends visual plots (Part 2 of 2)

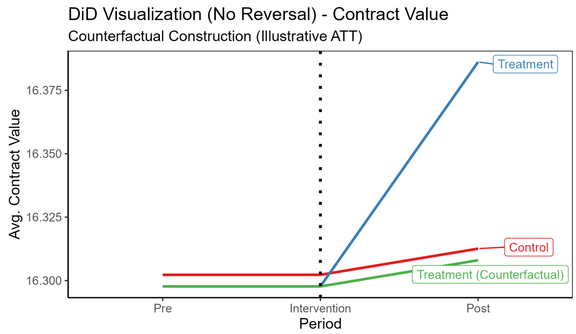
7.4 Appendix 4: Counterfactual Plots



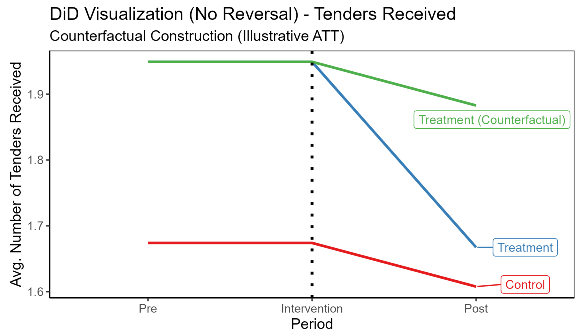
(4.1) Only Green in Post - Contract Value



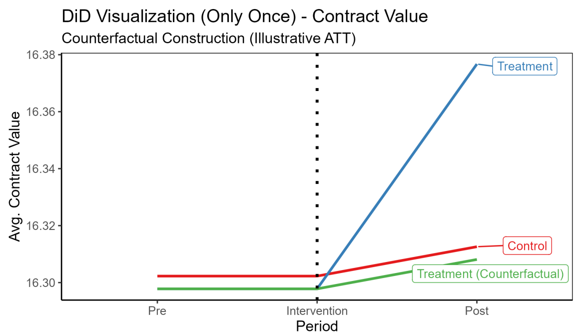
(4.2) Only Green in Post - Received Tenders



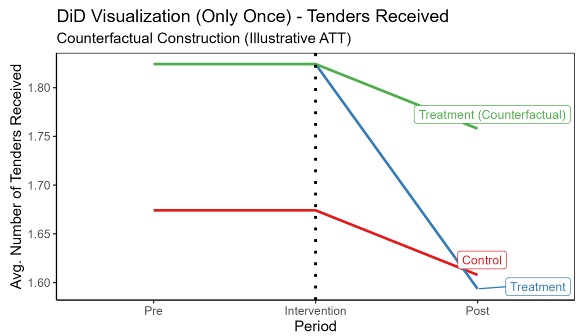
(4.3) No Reversal - Contract Value



(4.4) No Reversal - Received Tenders



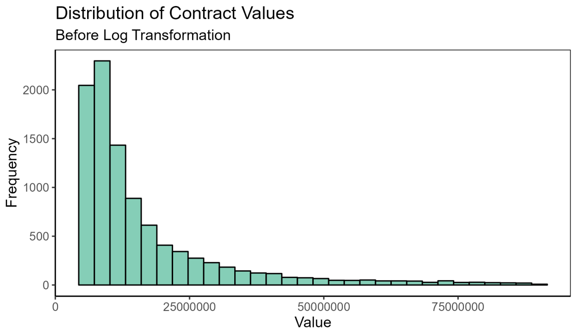
(4.5) Only Once - Contract Value



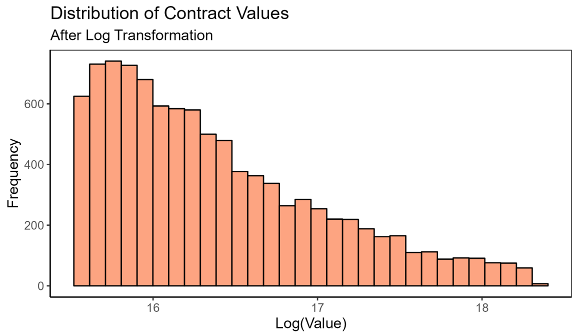
(4.6) Only Once - Received Tenders

Fig. 4: Counterfactual Trend Plots for all Configurations and Outcomes

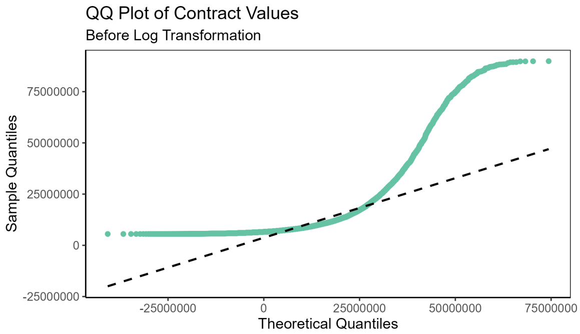
7.5 Appendix 5: Q-Q and Density Plots Pre- and Post Log Transformation



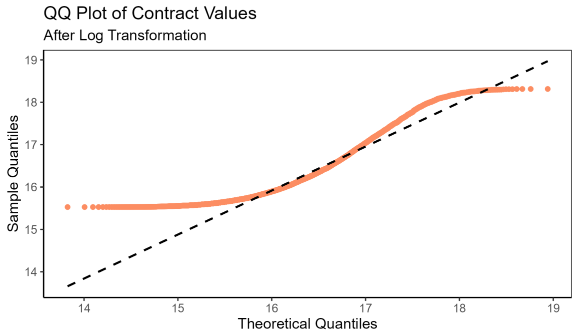
(5.1) Distribution of Contract Values (prior to log transformation)



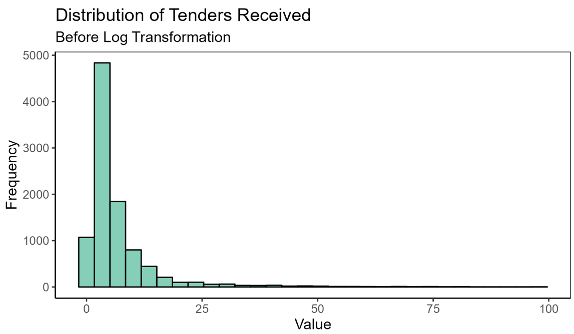
(5.2) Distribution of Contract Values (after log transformation)



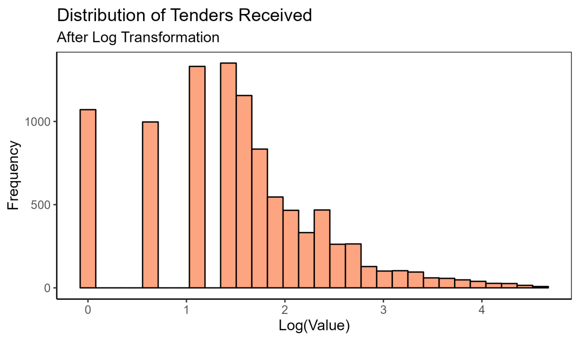
(5.3) Q-Q Plot Contract Values (prior to log transformation)



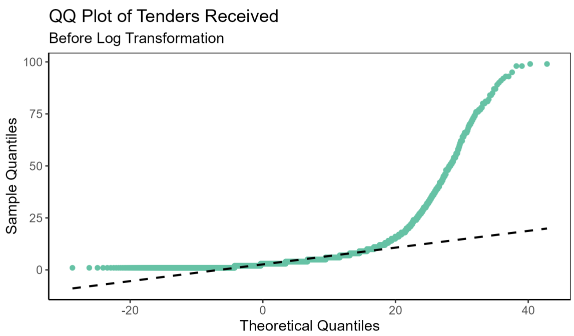
(5.4) Q-Q Plot Contract Values (after log transformation)



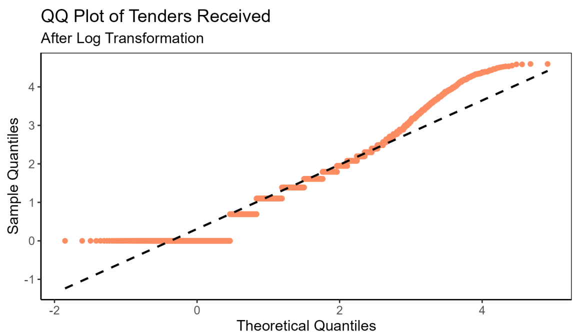
(5.5) Distribution of Received Tenders (prior to log transformation)



(5.6) Distribution of Received Tenders (after log transformation)



(5.7) Q-Q Plot Received Tenders (prior to log transformation)



(5.8) Q-Q Plot Received Tenders (after log transformation)

Fig. 5: Log Transformation Visualisations

7.6 Appendix 6: Results RQ1

Table 7: Utilization of Green Criteria Across Different Time Periods

<table>
<tr><th colspan="2">Year</th><th colspan="2">Total CFT</th><th colspan="2">% Green</th></tr>
<tr><td colspan="2">2012</td><td colspan="2">2195</td><td colspan="2">14.76</td></tr>
<tr><td colspan="2">2013</td><td colspan="2">2331</td><td colspan="2">13.99</td></tr>
<tr><td rowspan="2">2014</td><td>ex ante</td><td rowspan="2">2105</td><td>278</td><td rowspan="2">13.87</td><td>15.11</td></tr>
<tr><td>ex post</td><td>1827</td><td>13.68</td></tr>
<tr><td colspan="2">2015</td><td colspan="2">1952</td><td colspan="2">15.11</td></tr>
<tr><td rowspan="2">2016</td><td>ex ante</td><td rowspan="2">1736</td><td>467</td><td rowspan="2">18.26</td><td>14.56</td></tr>
<tr><td>ex post</td><td>1269</td><td>19.62</td></tr>
<tr><td colspan="2">2017</td><td colspan="2">2229</td><td colspan="2">25.44</td></tr>
<tr><td colspan="2">2018</td><td colspan="2">2379</td><td colspan="2">22.11</td></tr>
<tr><td colspan="2">2019</td><td colspan="2">2460</td><td colspan="2">24.15</td></tr>
<tr><td colspan="2">2020</td><td colspan="2">2477</td><td colspan="2">22.69</td></tr>
<tr><td colspan="2">2021</td><td colspan="2">2595</td><td colspan="2">27.09</td></tr>
<tr><td colspan="2">2022</td><td colspan="2">2739</td><td colspan="2">26.69</td></tr>
<tr><td colspan="2">2023</td><td colspan="2">2461</td><td colspan="2">31.33</td></tr>
<tr><td colspan="2">Pre</td><td colspan="2">4804</td><td colspan="2">14.40</td></tr>
<tr><td colspan="2">During</td><td colspan="2">4246</td><td colspan="2">14.44</td></tr>
<tr><td colspan="2">Post</td><td colspan="2">18609</td><td colspan="2">25.27</td></tr>
</table>

Table 8: Statistical Test Results for Proportions Across Different Time Periods

$Comparison$	χ^2	$p - value$	95% $Confidence\ Interval$
Pre vs. During	$1.9266e^{-4}$	0.9889	$[-0.0150 : +0.0144]$
Pre vs. Post	253.71	$2.2e^{-16}$ ***	$[-0.1205 : -0.0968]$
During vs. Post	228.49	$2.2e^{-16}$ ***	$[-0.1237 : -0.0981]$
Pre vs. 2014	0.50699	0.4764	$[-0.0118 : +0.0262]$
Pre vs. 2015	0.50248	0.4784	$[-0.0262 : +0.0120]$
Pre vs. 2016	14.235	$1.613e^{-4}$ ***	$[-0.0597 : -0.0175]$
2014 vs. 2015	1.4479	0.2289	$[-0.0372 : +0.0086]$
2014 vs. 2016	13.595	$2.268e^{-4}$ ***	$[-0.0704 : -0.0212]$
2015 vs. 2016	6.352	0.0117 **	$[-0.0562 : -0.0068]$
2016 (ante vs. post)	5.0401	0.0248 **	$[-0.0888 : -0.0081]$

Notes: $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$

7.7 Appendix 7: OLS Regression Results

Table 9: OLS Regression Results for Contract Values and Tenders Received Across Different Configurations

Coefficient	*Contract Value*			*Tenders Received*		
	Only Green	*No Reversal*	*Only Once*	*Only Green*	*No Reversal*	*Only Once*
Intercept	13.9810 *** (2.0684)	14.4021 *** (1.9922)	14.8018 *** (1.8452)	1.0845 (3.1782)	0.7232 (3.0883)	1.7875 (2.8489)
post1	−0.0534 (0.1038)	−0.0581 (0.0992)	−0.0425 (0.0908)	0.0886 (0.1548)	0.0388 (0.1573)	0.0865 (0.1402)
treat1	−0.1196 * (0.0611)	−0.0182 (0.0483)	−0.0173 (0.0933)	0.3131 *** (0.0984)	0.2465 *** (0.0749)	0.1389 ** (0.0607)
SDG	0.0407 (0.0265)	−0.0366 (0.0255)	0.0036 (0.0255)	0.0208 (0.0407)	0.0258 (0.0395)	0.0120 (0.0363)
GOVexp	−0.0661 (0.0103)	−0.0073 (0.0103)	−0.0345 (0.0910)	−0.0165 (0.0158)	−0.0186 (0.0156)	−0.0171 (0.0141)
covid	−0.3576 (0.8179)	−0.4063 (0.7995)	−0.3335 (0.7651)	−1.1505 (25.6565)	−1.1371 (1.2385)	−0.0189 (1.1813)
CPPI_growth	0.0023 (0.0157)	−0.0394 (0.0150)	0.0026 (0.0137)	0.0295 (0.2414)	−0.0227 (0.0233)	0.0279 (0.0212)
factor(CAE_TYPE_grouped)						
National	−0.7059 *** (0.2725)	−0.7026 *** (0.2702)	−0.7048 *** (0.2563)	−0.0227 (0.4187)	0.0164 (0.4183)	0.0430 (0.4096)
	2 levels suppressed					
Unspecified	−0.8099 *** (0.2711)	−0.8241 *** (0.2688)	−0.8037 *** (0.2642)	−0.0339 (0.4157)	−0.0324 (0.4167)	−0.0116 (0.4079)
factor(COUNTRY)						
Belgium	0.2092 (0.1663)	−0.1845 (0.1512)	−0.1186 (0.2538)	−0.1587 (0.4166)	−0.0596 (0.4240)	−0.0911 (0.4226)
	23 levels suppressed	***24 levels suppressed***		***23 levels suppressed***	***24 levels suppressed***	
Slovakia	0.5233 *** (0.1835)	0.4505 ** (0.1761)	0.4478 *** (0.1638)	−0.1381 (0.2819)	−0.1199 (0.2729)	−0.1241 (0.2529)
post1:treat1	0.1365 (0.0838)	0.0630 (0.0652)	0.0547 (0.0535)	−0.2503 * (0.1288)	−0.2284 ** (0.1011)	−0.1716 ** (0.0826)
RMSE	0.5118 (*on* 1469 *df*)	0.5081 (*on* 1600 *df*)	0.4999 (*on* 1792 *df*)	0.7865 (*on* 1469 *df*)	0.7876 (*on* 1600 *df*)	0.7718 (*on* 1792 *df*)
Multiple R²	0.0854	0.0813	0.0805	0.09248	0.0992	0.0931
Adjusted R²	0.0630	0.0601	0.0615	0.07024	0.0784	0.0744
F-statistic	3.812 *** (*on* 36 *and* 1469 *df*)	3.8270 *** (*on* 37 *and* 1600 *df*)	4.239 *** (*on* 37 *and* 1792 *df*)	4.158 *** (*on* 36 *and* 1469 *df*)	4.762 *** (*on* 37 *and* 1600 *df*)	4.971 *** (*on* 37 *and* 1792 *df*)

Notes: $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$. SE between parentheses.
The Central region and the International CAE type are the omitted reference categories.

7.8 Appendix 8: Variance Inflation Factors

Table 10: Variance Inflation Factors (GVIF and Adjusted GVIF) for each configuration

Variable	*Only Green*	*No Reversal*	*Only Once*
	GVIF		
post	15.4818	15.6012	15.1040
treat	2.2122	2.2603	2.2507
SDG	32.6474	32.0444	31.0549
GOVexp	25.1581	26.5762	25.3406
covid	1.2124	1.2100	1.2113
CPPI_growth	16.8314	16.8132	16.2342
factor(CAE_TYPE_grouped)	1.6575	1.6352	1.5939
factor(Country)	3181.4167	3353.4354	2924.8845
post:treat	2.2213	2.3004	2.4720
Spillover	145.5800	151.3308	140.2046
	Adjusted GVIF ($\mathtt{GVIF}^{1/(2\times\mathtt{Df})}$)		
post	3.9347	3.9498	3.8864
treat	1.4873	1.5034	1.5002
SDG	5.7138	5.6608	5.5727
GOVexp	5.0158	5.1552	5.0339
covid	1.1011	1.1000	1.1006
CPPI_growth	4.1026	4.1004	4.0292
factor(CAE_TYPE_grouped)	1.0652	1.0634	1.0600
factor(Country)	1.1750	1.1690	1.1659
post:treat	1.4903	1.5167	1.5722
Spillover	12.0657	12.3017	11.8408

Notes: All df = 1, except for CAE_TYPE_grouped (=4) and country (=25 in only green / = 26 others).

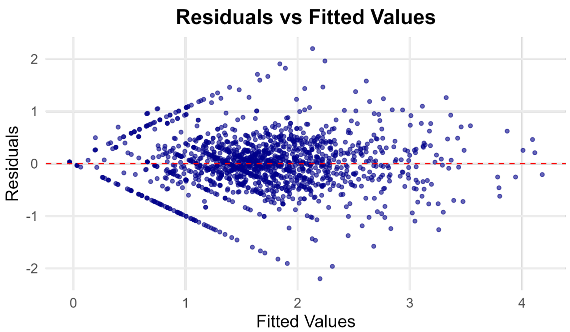
7.9 Appendix 9: RQ2 Regression Results

Table 11: One-Way Fixed Effects (CAE_ID) DiD Results for Tenders Received Across Different Configurations

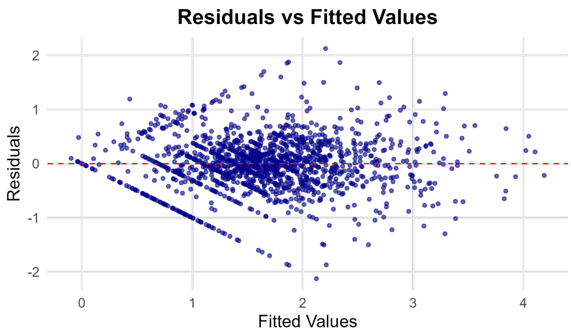
Coefficient	*Only Green*		*No Reversal*		*Only Once*	
	Baseline	*Adjusted*	*Baseline*	*Adjusted*	*Baseline*	*Adjusted*
post1	-0.0663^{*} (0.0379) $\|-1.751\|$	0.0387 (0.1784) $\|0.2168\|$	-0.0663^{*} (0.0379) $\|-1.7508\|$	0.0141 (0.1654) $\|0.0854\|$	-0.0663^{*} (0.0378) $\|-1.7510\|$	−0.0630 (0.1483) $\|0.4250\|$
SDG		0.0551 (0.0498) $\|1.1068\|$		0.0379 (0.0474) $\|0.7999\|$		0.0253 (0.0434) $\|0.5833\|$
GovExp		-0.0235^{*} (0.0132) $\|-1.7724\|$		-0.0238^{*} (0.0129) $\|-1.8417\|$		-0.0214^{*} (0.0117) $\|-1.8551\|$
Covid		−1.1262 (1.0235) $\|-1.1004\|$		−0.9682 (1.0166) $\|-0.9523\|$		−0.8194 (0.9908) $\|-0.8270\|$
CPPI_growth		−0.0291 (0.0294) $\|-0.9915\|$		−0.0217 (0.0270) $\|-0.8045\|$		−0.0273 (0.0239) $\|-1.1433\|$
post1:treat1	-0.2326^{**} (0.1114) $\|-2.0873\|$ [−.451 : −.014]	-0.2709^{**} (0.1204) $\|-2.2502\|$ [−.507 : −.035]	-0.2153^{***} (0.0826) $\|-2.6078\|$ [−.377 : −.053]	-0.2323^{***} (0.0881) $\|-2.6375\|$ [−.405 : −.059]	-0.1644^{**} (0.0656) $\|-2.5063\|$ [−.293 : −.036]	-0.1776^{**} (0.0707) $\|-2.5133\|$ [−.316 : −.039]
RMSE	0.4879	0.4860	0.4815	0.4801	0.4708	0.4693
Adjusted R²	0.2825	0.2841	0.3095	0.3101	0.3096	0.3108
Within R²	0.0148	0.0223	0.0195	0.0252	0.0196	0.0256
Panel Structure	N = 1506 T = 2 C = 753	N = 1506 T = 2 C = 753	N = 1638 T = 2 C = 819	N = 1638 T = 2 C = 819	N = 1830 T = 2 C = 915	N = 1830 T = 2 C = 915

Notes: $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$. SE between parentheses, t-value between straight lines, 95% conf. interval between square brackets. N represents the number of observations. T equals the number of time periods. C equals the number of clusters.

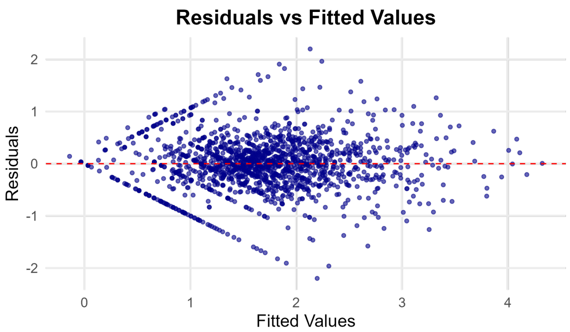
7.10 Appendix 10: RQ2 Residual Plots



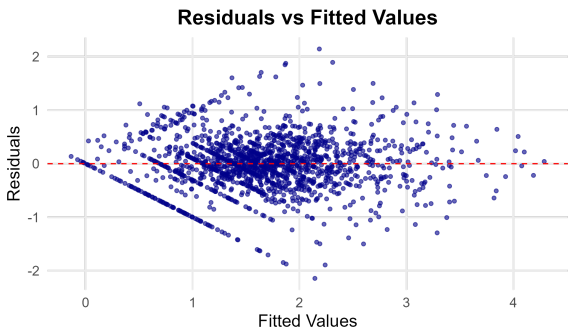
(6.1) Only Green Baseline - Residuals vs. Fitted Values - Received Tenders



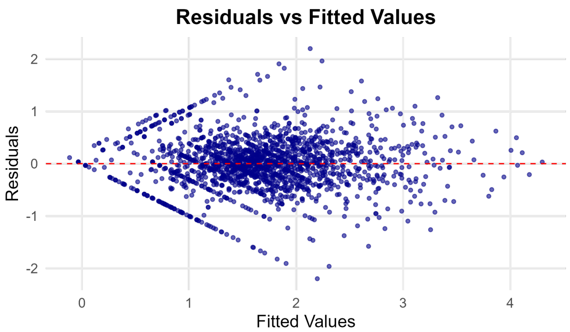
(6.2) Only Green Adjusted - Residuals vs. Fitted Values - Received Tenders



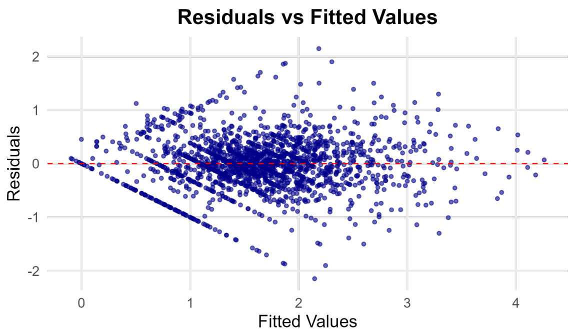
(6.3) No Reversal baseline - Residuals vs. Fitted Values - Received Tenders



(6.4) No Reversal baseline - Residuals vs. Fitted Values - Received Tenders



(6.5) Only Once baseline - Residuals vs. Fitted Values - Received Tenders



(6.6) Only Once baseline - Residuals vs. Fitted Values - Received Tenders

Fig. 6: Residual vs. Fitted Values Plots for all Received Tenders Baseline and Adjusted Models

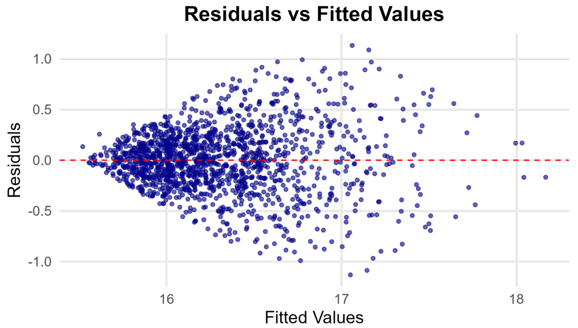
7.11 Appendix 11: RQ3 Regression Results

Table 12: One-Way Fixed Effects (CAE_ID) DiD Results for Contract Values Across Different Configurations

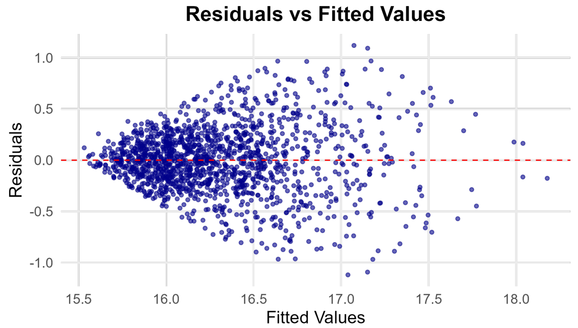
Coefficient	*Only Green*		*No Reversal*		*Only Once*	
	Baseline	*Adjusted*	*Baseline*	*Adjusted*	*Baseline*	*Adjusted*
post1	0.0103 (0.0245) |0.4213|	−0.0414 (0.0970) |−0.4265|	0.0103 (0.0245) |0.4213|	−0.0560 (0.0930) |−0.6020|	0.0103 (0.0245) |0.4213|	−0.0414 (0.0839) |−0.4937|
SDG		0.0389 (0.0301) |1.2876|		0.0425 (0.0287) |1.4788|		0.0397 (0.0267) |1.4860|
GovExp		−0.0027 (0.0092) |−0.2939|		−0.0047 (0.0090) |−0.5255|		−0.0019 (0.0082) |−0.2349|
Covid		−0.8913 (0.7568) |−1.1777|		−1.1456 (0.7343) |−1.5600|		−1.0613 (0.7234) |−1.4671|
CPPI_growth		0.0017 (0.0145) |0.1163|		0.0036 (0.0140) |0.2553|		0.0018 (0.0124) |0.1446|
post1:treat1	0.1620 ** (0.0801) |2.0214| [.0047 : .3193]	0.1362 * (0.0826) |1.6493| [−.0259 : .2982]	0.0781 (0.0582) |1.3425| [−.0361 : .1924]	0.0584 (0.0592) |0.9858| [−.0579 : .1746]	0.0686 (0.0465) |1.4765| [−.0226 : .1598]	0.0507 (0.0474) |1.0698| [−.0423 : .1436]
RMSE	0.3204	0.3199	0.3175	0.3169	0.3143	0.3137
Adjusted R²	0.2638	0.2619	0.2642	0.2634	0.2565	0.2557
Within R²	0.0086	0.0114	0.0038	0.0076	0.0045	0.0078
Panel Structure	N = 1506 T = 2 C = 753	N = 1506 T = 2 C = 753	N = 1638 T = 2 C = 819	N = 1638 T = 2 C = 819	N = 1830 T = 2 C = 915	N = 1830 T = 2 C = 915

Notes: $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$. SE between parentheses, t-value between straight lines, 95% conf. interval between square brackets. N represents the number of observations. T equals the number of time periods. C equals the number of clusters.

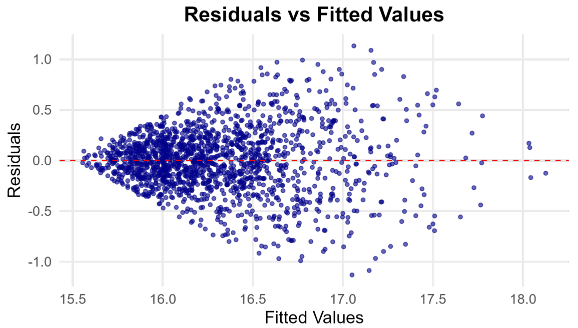
7.12 Appendix 12: RQ3 Residual Plots



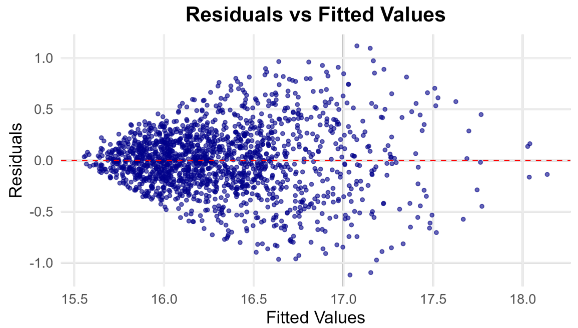
(7.1) Only Green Baseline - Residuals vs. Fitted Values - Contract Value



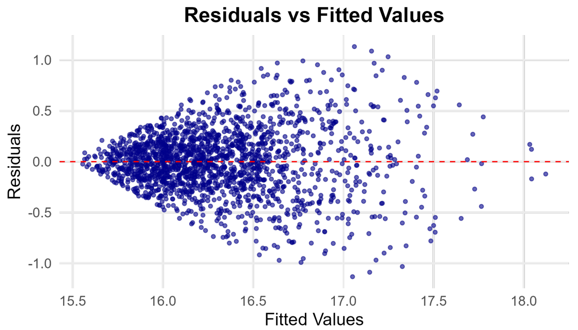
(7.2) Only Green Adjusted - Residuals vs. Fitted Values - Contract Value



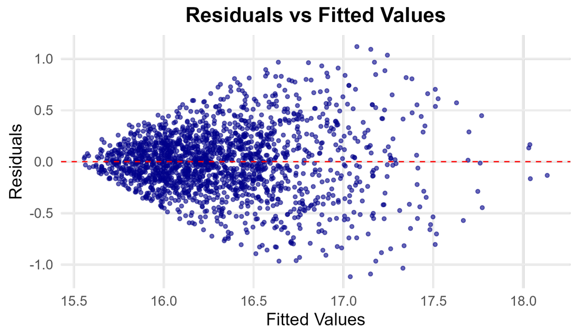
(7.3) No Reversal baseline - Residuals vs. Fitted Values - Contract Value



(7.4) No Reversal baseline - Residuals vs. Fitted Values - Contract Value



(7.5) Only Once baseline - Residuals vs. Fitted Values - Contract Value



(7.6) Only Once baseline - Residuals vs. Fitted Values - Contract Value

Fig. 7: Residual vs. Fitted Values Plots for all Contract Value Baseline and Adjusted Models

7.13 Appendix 13: Aggregation Structure Visualized

Table 13: Overview of procurement structure at CN and CAN level, and the applied aggregation approach

Posted in CN Dataset				*(Announced tenders)*
Contracting Authority	CN ID	CAN ID	Lot ID	Contract Award
X	1	–	1	–
X	1	–	2	–
X	2	–	1	–
X	3	–	1	–
X	3	–	2	–
Posted in CAN Dataset				*(Awarded results)*
Contracting Authority	CN ID	CAN ID	Lot ID	Contract Award
X	1	1	1	1
X	1	1	2	2
X	3	2	1	1
X	3	2	2	2 (duplicate)
X	3	2	TWO	2 (duplicate)
Aggregated View (Applied Approach)				*(Grouped by CAN)*
Contracting Authority	CN ID	CAN ID	Lot ID	Aggregated CAs
X	1	1	–	1;2
X	3	2	–	1;2

Notes: CN = Contract Notice (initial tender); CAN = Contract Award Notice (award result); CA = Contract Award. In the aggregated view, multiple contract awards belonging to the same CAN are grouped to mitigate issues with unstructured lot identifiers and duplicate entries.

7.14 Appendix 14: PSM-Matching Statistics

Pre-Pre match

```
Summary of Balance for All Data:
                             Means Treated Means Control Std. Mean Diff. Var. Ratio eCDF Mean eCDF Max
distance                            0.1530        0.0794          1.0642     0.8870    0.2651   0.4186
Summary of Balance for Matched Data:
                             Means Treated Means Control Std. Mean Diff. Var. Ratio eCDF Mean eCDF Max Std. Pair Dist.
distance                            0.1530        0.1530          0.0005     1.0032    0.0001   0.0040           0.0010
Sample Sizes:
          Control Treated
All         10759    1008
Matched      1008    1008
Unmatched    9751       0
Discarded       0       0
```

Pre-Post match

```
Summary of Balance for All Data:
                             Means Treated Means Control Std. Mean Diff. Var. Ratio eCDF Mean eCDF Max
distance                            0.6560        0.1131          1.4813     8.6434    0.3679   0.6390
Summary of Balance for Matched Data:
                             Means Treated Means Control Std. Mean Diff. Var. Ratio eCDF Mean eCDF Max Std. Pair Dist.
distance                            0.6560        0.2452          1.1209     7.3250    0.1443   0.5785           1.1209
Sample Sizes:
          Control Treated
All         10759    3537
Matched      3537    3537
Unmatched    7222       0
Discarded       0       0
```

Fig. 8: Summary of Covariate Balance Statistics (Pre-Pre & Pre-Post)

7.15 Appendix 15: List of Environmental Keywords

This appendix contains the complete list of keywords used to identify environmental award criteria in procurement notices. The terms were compiled based on relevant standards, legislation, and recurring terminology identified in sustainable procurement literature. While some terms may appear general (e.g., 'certification', 'maintenance'), they were included due to their frequent use in environmentally focused award criteria.

- 13790
- 14000
- 14001
- 14024
- 14025
- 15251
- 15804
- 50001
- AECI
- BIM
- BREEAM
- CF
- CO2
- COD
- Circular Economy Action Plan
- Climate Action Regulation
- Construction Products Regulation
- Cradle to Cradle
- Directive 2014/24/EU
- EIA
- EMAS
- EMS
- EN 15251
- EN 15804
- EPC
- EPD
- Energy Efficiency Directive
- Energy Performance of Buildings Directive
- FSC
- GHG
- GWP
- Green
- Green Deal
- Green Public Procurement
- ISO 13790
- ISO 14000
- ISO 14001
- ISO 14024
- ISO 14025
- ISO 50001
- LCA
- LCC
- LEED
- MKI
- NDC
- NOX
- NZEB
- OGC
- PEFC
- PM
- PM10
- PM2.5
- RES
- SEA
- SUDS
- Sulfur emissions
- TiO
- VOC-free paints
- Waste Framework Directive
- Zero emission
- Zero-energy buildings
- air pollution
- air quality
- air quality thresholds
- biodegradable
- biodiversity protection
- carbon
- carbon accounting
- carbon footprint
- carbon intensity
- certificates
- certification
- certified
- chemical restriction
- circular
- circular economy
- climate neutrality
- climate resilience analysis
- dimming
- directive 2014/24/eu
- durability
- durable
- eco
- eco-design
- eco-efficiency analysis
- eco-friendly
- eco-label
- ecolabel
- ecological
- ecology
- emission
- emission reduction targets
- emissionless
- energy
- energy assesment
- energy balance calculation
- energy consumption
- energy efficiency
- energy modeling
- energy performance
- energy performance certificate
- energy saving
- environment
- environmental
- environmental cost-benefit analysis
- environmental criteria
- environmental footprint
- environmental impact
- environmental performance evaluation
- fossil-free
- fuel consumption
- gas consumption
- geothermal systems
- green building
- green procurement
- greenhouse gas
- hazard labeling
- hazardous
- heat recovery systems
- innovation in design
- insulation standards
- lifecycle
- life cycle
- life cycle assessment
- life cycle cost
- lifecycle cost
- lifespan
- lifetime
- low-emission
- low-impact construction materials
- low-noise
- maintenance
- management of waste
- material flow analysis
- maximum CO_2 emissions threshold
- net-zero carbon
- noise emissions
- noise pollution limits
- non-toxic materials
- organic
- organic gaseous carbon
- particulate matter
- passive design
- percentage of recycled content
- photovoltaic panels
- pollutant
- pollution
- pollution control
- rainwater harvesting systems
- re-use
- rechargeable
- recover fibers
- recyclable
- recycle
- recycled
- recycled materials
- recycling
- reduction in embodied carbon
- reduction of light energy consumption
- renewable
- renewable energy
- renewable energy directive
- renewable energy generation capacity
- repair
- repairability
- repairable
- reparability
- replacement
- resource efficiency
- reusable
- reuse
- reused
- soil management
- sustainability
- sustainable
- sustainable construction
- sustainable material sourcing
- sustainable materials
- sustainable mobility infrastructure
- sustainable procurement
- sustainable urban development
- sustainable urban drainage
- thermal insulation U-value
- titanium dioxide
- waste
- waste conservation
- waste consumption per unit
- waste diversion rate
- waste management
- waste management plan
- waste prevention
- waste recovery
- waste reduction
- water conservation
- water consumption
- water efficiency
- water footprint analysis
- water saving
- whole life cost
- wildlife

7.16 Appendix 16: Tendering Cycle: Open Procedure

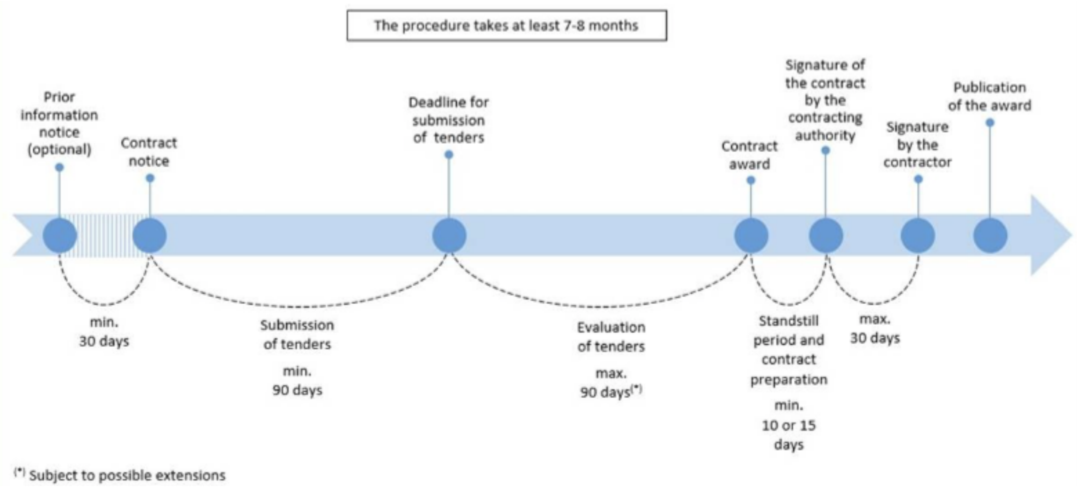


Fig. 9: Timeline Open Tender Procedures for Works Contracts. Reprinted from Contract Procedures for EU External Action: A practical guide, by European Commission [21].