Did Spain's Action Plan Against Hate Crimes Increase Reporting? Evidence from a Causal Impact Analysis

Alberto Matilla-Molina

amatilla@interior.es

Andrés Muñoz

andres.munoz@uca.es

Abstract

Hate crime underreporting in Spain constrains victim support and evidence-based policy. We test whether Spain's Action Plan to Combat Hate Crimes (2019-2021) increased recorded incidents by improving reporting and detection. Using Bayesian structural time series causal impact analysis, we model monthly counts from Spain's Statistical System of Criminality (January 2015-December 2021) with Germany and Finland as predictive covariates. The model learns the pre-intervention relationship and forecasts the post-intervention counterfactual. Recorded incidents increased by 20.3% (95% credible interval 10.5%-30.3%); excluding 2020, the increase was 27.3% (16.3%-39.2%). Placebo and sensitivity tests support a policy-specific effect (posterior probability > 0.99). Findings are consistent with improved awareness, recording, and reporting rather than higher victimisation.

Keywords

hate crimes, underreporting, policy evaluation, Spain, causal inference.

1. Introduction

Hate crimes represent serious violations of human rights that extend beyond individual victims to undermine social cohesion and challenge fundamental principles of equity in democratic societies (Herek, 2009; Jacobs & Potter, 1998). Despite their severity, hate crimes are drastically underreported. In the European Union Minorities and Discrimination Survey (EU-MIDIS), conducted by the European Union Agency for Fundamental Rights (FRA), 75%-90% of serious harassment incidents were not reported to the police, depending on the group and context (FRA, 2010). More recent FRA evidence continues to show very low reporting levels for bias-motivated harassment, indicating that underreporting remains pervasive across the EU (FRA, 2021). This persistent gap between actual victimization and official records poses a critical challenge for policymakers: interventions cannot effectively address a problem that remains largely invisible in official statistics.

Spain exemplifies this challenge acutely. With an estimated 90% underreporting rate (López Gutiérrez et al., 2021), official records captured only 12,983 hate crime incidents between 2013 and 2021 (Ministry of Interior, 2023). Recognizing this gap, the Spanish Ministry of Interior implemented the "Action Plan to Combat Hate Crimes" between 2019 and 2021. This comprehensive intervention encompassed 54 measures focused on training security forces, preventing hate incidents, supporting victims, and strengthening institutional responses (Ministry of Interior, 2019). A central goal was improving detection and reporting mechanisms to reduce the dark figure of unreported hate crimes.

The meaning of these official counts depends on how Spanish law and institutions define and record hate-related conduct; Section 2.1 sets out the conceptual and legal framing used throughout the article.

Despite the proliferation of hate crime policies across Europe, rigorous causal evaluations of their effectiveness remain scarce. Existing literature has advanced understanding of hate crime dynamics through time-series forecasting (Dorsett, 2021), computational approaches combining machine-learning-based hate speech detection with police-recorded crime data (Williams et al., 2020), and analyses linking hate crimes to political events such as Brexit or immigration influxes (Piatkowska & Lantz, 2021; Piatkowska et al., 2020). However, most policy assessments rely on descriptive or correlational evidence, leaving a critical question unanswered: do government interventions genuinely alter hate crime reporting and incidence, or do observed changes reflect broader social trends unrelated to policy implementation?

This study addresses this gap by applying causal inference methods to evaluate Spain's Action Plan. Specifically, we employ Bayesian Structural Time Series (BSTS) modeling to estimate the policy's causal impact on recorded hate crimes, using data from Germany and Finland as comparative controls to isolate Spain-specific effects from broader European trends. Accordingly, we expect the Action Plan to increase officially recorded hate-related incidents primarily by improving reporting, detection, and recording mechanisms, rather than by generating a genuine rise in hate-motivated victimisation. This distinction carries important policy implications: an increase in reporting signals successful intervention, whereas an increase in actual incidents would indicate policy failure.

This research contributes to both scholarship and practice in three ways. First, it provides the first causal impact evaluation of a national hate crime policy in Spain, offering empirical evidence where previously only descriptive assessments existed. Second, it demonstrates the utility of integrating international comparative data in policy evaluation, a methodological approach applicable to other contexts characterized by

significant underreporting. Third, it offers actionable insights for policymakers and law enforcement agencies seeking evidence-based strategies to combat hate crimes-insights particularly relevant as European nations strengthen their responses to bias-motivated offenses.

2. Background

2.1 Conceptual and legal framing of hate crime in Spain

Although the term hate crime is widely used in comparative scholarship, the Spanish framework requires greater conceptual precision. For the purposes of this article, hate crime is understood analytically as criminal or institutionally relevant conduct in which prejudice or discriminatory bias against a protected group, or against a person perceived as belonging to such a group, plays a constitutive role in the event. In Spain, however, this phenomenon is not exhausted by a single offence formally labelled as "hate crime." Instead, the legal response is distributed across a broader architecture that combines the generic aggravating circumstance for discriminatory motivation in Article 22.4 of the Criminal Code with a range of specific offences, most notably Article 510, which regulates serious forms of public incitement to hatred, discrimination, hostility, or violence. The protected grounds recognised in the Spanish criminal-law framework include, among others, ideology, religion or beliefs, ethnicity, race, nation, sex, age, sexual orientation or identity, gender, aporophobia or social exclusion, illness, and disability. In parallel, institutional monitoring and statistical reporting refer to delitos e incidentes de odio, thereby bringing together a broader set of hate-related behaviours as they are identified and recorded by the competent authorities (Ley Orgánica 10/1995; Fiscalía General del Estado, 2019; Ministry of Interior, 2024).

This distinction is analytically important for the present study. Official Spanish data do not represent a single, perfectly self-contained legal category; rather, they reflect the operation of a broader institutional framework through which hate-related conduct becomes visible to the state. For that reason, this article uses the expression recorded hate-related incidents when referring to the outcome under study, while retaining the more familiar term hate crime where needed for readability and international comparison. This clarification is especially important in cross-national research, since comparative work has shown that hate crime, hate incidents, and hate speech are defined and measured unevenly across jurisdictions, which directly affects prevalence estimates, comparability, and policy evaluation (FRA, 2018; Vergani et al., 2024).

2.2 Determinants of Hate Crime Incidence and Reporting

Understanding hate crime patterns requires examining both the factors that drive incidents and those that influence whether victims report them. Research has consistently identified sociopolitical triggers that precipitate spikes in hate offenses. Large-scale immigration events activate perceived group threat among local populations, correlating with increased right-wing hate crimes (Piatkowska et al., 2020). Similarly, major political events-such as the Brexit referendum and terrorist attacks-generate temporal clustering of hate offenses, with intensity varying according to local political climates and public discourse on immigration (Piatkowska & Lantz, 2021; Piatkowska & Stults, 2022). These findings underscore that hate crime rates respond dynamically to broader social and political contexts, complicating efforts to isolate policy effects from environmental influences.

Reporting behavior presents equally complex dynamics. Victims' decisions to report hate incidents depend on trust in authorities, perceived institutional responsiveness,

procedural accessibility, and awareness of reporting mechanisms (FRA, 2018, 2021). The 2024 Spanish Victim and Witness Survey revealed that most respondents who experienced repeated hate incidents chose not to report them, citing mistrust in authorities and perceived ineffectiveness of institutional responses (Muniesa Tomás et al., 2024), and Wiedlitzka et al. (2018) show that more positive perceptions of police legitimacy are associated with victims' decisions to report hate crime. This evidence suggests that official statistics reflect not only actual crime rates but also victims' willingness to engage with the criminal justice system—and provides a clear theoretical basis for expecting institutional interventions to affect recorded counts even when underlying incidence is stable.

2.3 Spain's Action Plan to Combat Hate Crimes: content and expected mechanisms

Spain's Action Plan to Combat Hate Crimes, in force from 2019 to 2021, was conceived as a comprehensive public-policy package rather than as a single operational measure. According to the Ministry of the Interior, it comprised 54 measures aimed at strengthening police training, preventing hate incidents, improving victim support, and reinforcing institutional responses. Subsequent official reporting indicated that 48 of the 54 measures had been materialised by the end of the period, suggesting a substantial degree of implementation (Ministry of Interior, 2019, 2022).

Several of these measures took concrete form during the implementation window. In the training axis, specialised instruction on the identification of bias indicators was incorporated into the training plans of the National Police and Civil Guard, and—as one of the Plan's own measures—the operational Protocol for Law Enforcement Action on Hate Crimes was comprehensively updated in July 2020 (Instruction 4/2020 of the Secretariat of State for Security), including an expanded catalogue of criminal and

administrative hate-related conducts and a detailed list of "polarisation indicators" to guide officers in detecting and recording bias motivation (Ministry of Interior, 2020). In the awareness axis, the Plan mandated institutional campaigns by the National Police, intensified around commemorative dates and in response to serious incidents, alongside structured cooperation with civil society organisations working with frequently targeted groups. In the victim-support axis, the figure of the police "social interlocutor" for hate crimes was deployed at national and territorial levels as a stable liaison between law enforcement and vulnerable communities, and reporting was made more accessible through a dedicated hate-crime functionality added in 2020 to Alertcops, the Ministry's citizen security mobile application, allowing victims and witnesses to alert police and transmit evidence telematically. These examples do not exhaust the 54 measures, but they illustrate how each mechanism was operationalised in practice.

From an evaluative perspective, these measures can be grouped into four mechanisms that are directly relevant to reporting outcomes. First, professionalisation and training can increase police capacity to recognise bias indicators and classify incidents more consistently. Second, awareness and prevention measures can improve victims' and witnesses' understanding of what qualifies as a reportable hate-related incident. Third, victim-support measures can lower the practical and psychological costs of engaging with authorities. Fourth, coordination and institutional-response measures can improve the visibility, standardisation, and continuity of recording practices across agencies. These mechanisms are well aligned with the barriers identified in EU-level work on hate crime reporting and recording, which highlights the importance of accessible reporting routes, trained personnel, and coordinated institutional responses (FRA, 2018, 2021).

This mechanism-based reading is critical for interpreting the empirical results. The Action Plan did not target only primary prevention of offending; it also sought to improve detection, reporting, and institutional response. For that reason, an increase in recorded incidents after implementation is compatible with policy success in reducing the dark figure of hate crime, even though it does not by itself prove a change in underlying victimisation.

2.4 Policy Responses and Their Evaluation

Governments across Europe have implemented diverse strategies to address hate crimes, ranging from legislative reforms to awareness campaigns and specialized police training. Johnson (2023) demonstrated that the clarity of written hate crime guidelines significantly enhances law enforcement's efficacy in identifying and handling incidents, while Chai (2025) identified institutional barriers that hinder policy implementation within organizational hierarchies. At the macro level, Farrell and Lockwood (2023) and Dancygier (2023) have documented how policy changes interact with cultural shifts to shape patterns of bias-motivated offenses.

Despite this growing body of work, rigorous causal evaluations of hate crime policies remain notably absent. Most assessments rely on pre-post comparisons without adequate controls for confounding trends, or on correlational analyses that cannot establish causality. Methodological advances in adjacent fields offer promising tools: Bayesian structural time series models have successfully evaluated policy interventions in contexts ranging from crime reduction programs (Clark-Moorman et al., 2019) to economic policies (Dorsett, 2021), while difference-in-differences designs have become standard for causal inference in policy evaluation (Card & Krueger, 2000; Callaway &

Sant'Anna, 2021). Yet these methods have rarely been applied to hate crime interventions, leaving policymakers without robust evidence of what works.

These considerations also clarify the broader relevance of the Spanish case beyond Spain and beyond Europe. In contexts marked by substantial underreporting, the evaluation of hate crime policy cannot be limited to asking whether incidence falls. It must also ask whether the state becomes better able to recognise, record, and respond to bias-motivated harm. That problem is not uniquely Spanish. Comparative reviews continue to emphasise both the social harm of hate crime and the importance of improving reporting, recording, victim support, and institutional coordination across jurisdictions (Farrell & Lockwood, 2023; FRA, 2021). The present study therefore speaks not only to one national action plan, but also to a broader international problem of how to evaluate policy effectiveness when official counts reflect both victimisation and institutional visibility.

To address this methodological gap, we apply Bayesian Structural Time Series (BSTS) causal inference to evaluate Spain's Action Plan to Combat Hate Crimes (2019-2021). Our approach differs from prior work in three respects: (i) we use a counterfactual framework that estimates what recorded hate incidents would have looked like in the absence of the intervention, rather than relying on simple pre-post comparisons; (ii) we incorporate international time series (Germany and Finland) as predictive covariates to improve pre-intervention fit and counterfactual estimation; and (iii) we report multiple robustness and falsification checks (including placebo-period tests and sensitivity analyses) to strengthen causal interpretation.

3. Methodology

3.1 Analytic Approach: Bayesian Structural Time Series

Evaluating the causal impact of policy interventions requires comparing observed outcomes to a counterfactual scenario representing what would have occurred absent the intervention. While randomized experiments provide the gold standard for such comparisons, they are rarely feasible in public policy contexts (Angrist & Pischke, 2009). This study employs Bayesian Structural Time Series (BSTS) modeling, a quasi-experimental approach that constructs counterfactual predictions using pre-intervention data and external control series unaffected by the treatment.

We implement the BSTS framework through Google's CausalImpact model (Brodersen et al., 2015), which has been successfully applied to evaluate interventions in criminology (Clark-Moorman et al., 2019), public health, and economic policy (Dorsett, 2021). The approach offers three key advantages for our research question. First, it accommodates non-stationary trends common in crime data, avoiding the stationarity assumptions required by traditional time-series methods. Second, it incorporates external covariates-in our case, hate crime trends from Germany and Finland-to improve counterfactual predictions. Third, it provides full posterior distributions of causal effects, enabling probabilistic statements about intervention impacts rather than simple point estimates.

The core logic is straightforward: the model learns the relationship between Spanish hate crime trends and the control series during the pre-intervention period, then projects what Spanish trends would have looked like post-intervention based on this learned relationship. The difference between observed and predicted values constitutes the estimated causal effect.

3.2 Data and Measures

This section describes the variables, time window, and data provenance used for counterfactual estimation. The dependent variable is the monthly number of recorded hate-related incidents in Spain, drawn from the Statistical System of Criminality (SEC/SSC) for January 2015 to December 2021 (84 observations). In the Spanish statistical system, these are "hechos conocidos" by law-enforcement bodies, that is, incidents that come to police attention through complaints, police action, or other formal channels and are then incorporated into national crime statistics. Official Ministry reports on hate crime expressly present delitos e incidentes de odio registered by the Fuerzas y Cuerpos de Seguridad and reported to the SEC (Ministry of Interior, 2022, 2024).

This outcome should not be interpreted as a direct measure of total hate-crime victimisation. Rather, it reflects a composite institutional process involving victim reporting, police detection, recording practices, and legal or administrative classification. Spanish hate-crime monitoring also covers hate speech and related hate incidents within the official reporting framework, which makes the series well suited to evaluating changes in official visibility and registration, but less suited to isolating underlying victimisation on its own (Fiscalía General del Estado, 2019; Ministry of Interior, 2024; Vergani et al., 2024).

For readability, we use the conventional term *hate crime* throughout the article to refer to the broader phenomenon defined in Section 2.1. Analytically, however, the dependent variable and estimand in this study concern officially recorded hate-related incidents under the Spanish statistical and institutional framework, rather than a single self-contained offence category.

The control covariates are monthly police-registered hate-crime totals from Germany and Finland for the same period. These countries were selected because

continuous official monthly series were available for the full study window and because both jurisdictions participate in structured European hate-crime monitoring and reporting frameworks. At the same time, they are not treated as fully harmonised legal equivalents of Spain: Germany records hate crime within the broader system of politically motivated crime, while Finland's police statistics are compiled through a dedicated suspected hate-crime monitoring process (OSCE/ODIHR, n.d.-a, n.d.-b; Ministry of the Interior of Finland, n.d.). Their role in the model is predictive: the identifying assumption is not that Germany and Finland are legally or institutionally comparable to Spain, but that, absent the intervention, their series would have tracked the Spanish outcome in a sufficiently informative way—that is, that the statistical relationship between the three series learned over the pre-intervention period would have persisted. Two pieces of evidence support this assumption empirically. First, the pre-intervention model reproduces over half of the variation in the Spanish series ($R^2 = 0.55$; Table 1), with stable performance when either covariate is dropped. Second, placebo intervention windows located entirely within the pre-period show no spurious divergence (Section 4.2), indicating that the learned relationship does not break down arbitrarily. Like any extrapolation, the assumption could fail if either control country experienced a country-specific shock to hate-crime recording coinciding with Spain's intervention window; we return to this in Section 5.3.

The Action Plan was implemented from January 2019 to December 2021. We define the pre-intervention period as January 2015 to December 2018 (48 months) and the post-intervention period as January 2019 to December 2021 (36 months). Because 2020 may have affected both incidence and reporting, we also report parallel analyses excluding that year. Germany's monthly series was provided by the German Federal Ministry of the Interior via official correspondence as an aggregated monthly total, and Finland's monthly series was transmitted via the Finnish Ministry of Justice after retrieval

from the Police University College, also as aggregated monthly totals without bias-type or offence-type breakdowns.

3.3 Identification assumptions and pre-intervention fit

To interpret the estimated post-intervention deviations as a causal effect, two identification assumptions are required. First, the covariate series used to construct the counterfactual (Germany and Finland) are assumed not to be affected by the Spanish Action Plan itself (covariate exogeneity with respect to the intervention). Second, the statistical relationship between Spain's outcome series and the covariates is assumed to be stable over the pre-intervention period, such that it can be extrapolated into the post-intervention period absent the policy.

The plausibility of the second assumption for our specific covariates—Germany and Finland—was discussed in Section 3.2 and is assessed empirically in Table 1 and in the placebo-window tests of Section 4.2.

In our design, the intervention date is set to January 2019 because this is the official start of the Action Plan (i.e., when the policy is considered to begin operating). To provide transparent evidence that the covariates capture the pre-intervention dynamics of the Spanish series, we report pre-intervention fit diagnostics based on a simple model that includes (i) the selected covariates, (ii) a linear time trend, and (iii) monthly fixed effects to account for seasonality. Table 1 summarizes the pre-intervention performance for the baseline covariate set (Germany + Finland), as well as leave-one-covariate-out specifications. Across specifications, the model reproduces a substantial share of the pre-intervention variation and yields high correlations between observed and fitted values, supporting the plausibility of the counterfactual constructed by the main BSTS model.

As an additional sensitivity check regarding the intervention timing, we also evaluate pre-intervention fit under an alternative intervention date (July 2019), which expands the pre-intervention window through June 2019. This check is not intended to redefine the policy start date (which is fixed at January 2019 by design), but rather to verify that the pre-intervention relationship remains broadly consistent when the pre-period definition is varied.

Table 1. Pre-intervention fit diagnostics (Spain outcome).

Specification	Pre period (months)	RMSE	MAPE (%)	Correlation (r)	R^2
Baseline covariates: Germany + Finland; trend + month fixed effects	48	18.97	12.57	0.745	0.555
Leave-one-out: Germany only; trend + month fixed effects	48	19.11	12.44	0.740	0.548
Leave-one-out: Finland only; trend + month fixed effects	48	20.07	13.52	0.708	0.501
Alternative intervention timing: pre through 2019-06 (intervention 2019-07); Germany + Finland; trend + month fixed effects	54	19.95	13.68	0.720	0.518

Note: RMSE = root mean squared error; MAPE = mean absolute percentage error. Correlation is Pearson's r between observed and fitted values in the pre-intervention period. R2 is computed on the pre-intervention sample. These diagnostics document pre-intervention tracking and do not substitute for the Bayesian BSTS/CausalImpact estimation used for the main effect sizes.

3.4 Model Specification

The BSTS model decomposes the observed hate crime series into structural components:

$y_t = \mu_t + \gamma_t + \beta' X_t + \varepsilon_t$— *(Equation 1)*

where:

- y_t = observed hate crimes in Spain at time t
- μ_t = trend component capturing long-term trajectory
- γ_t = seasonal component capturing regular temporal patterns
- $\beta' X_t$ = regression component incorporating German and Finnish hate crime data as covariates
- ε_t = error term, assumed normally distributed

The model estimates latent states using the Kalman filter and samples from posterior distributions via Markov Chain Monte Carlo (MCMC) methods. These technical details, while essential for implementation, are standard in BSTS applications; readers interested in the mathematical foundations are referred to Brodersen et al. (2015) and the Causal Impact documentation. The causal effect at each time point is estimated as:

$Causal\ Effect = y_t - \hat{y}_t$— *(Equation 2)*

where y_t is the observed value and $\hat{y}_t$ is the counterfactual prediction. We report both average and cumulative effects across the post-intervention period, along with 95% credible intervals derived from the posterior distribution.

3.5 Robustness and Validation Tests

To ensure that the observed effects were not driven by random fluctuation or model artefacts, we conducted four validation procedures following established guidelines for causal inference (Imbens & Rubin, 2015; Pearl, 2009). First, we ran a placebo test using a standard normal outcome unrelated to the intervention; a valid model should detect no meaningful effect in this setting. Second, we added a synthetic random covariate to assess sensitivity to potential unobserved confounding; stability of the estimated effect under this perturbation supports robustness. Third, we re-estimated the model on a randomly selected 75% subset of observations using a fixed seed; consistency across subsets supports generalisability. Fourth, we implemented two placebo intervention windows entirely within the pre-intervention period; null findings in those windows support the interpretation that the observed post-2018 divergence is policy-specific rather than an artefact of pre-existing temporal structure. Complementing these checks, Table 1 reports pre-intervention tracking diagnostics and leave-one-covariate-out sensitivity analyses to document the stability of the pre-period fit.

4. Results

4.1 Main Findings

The BSTS analysis indicates an estimated increase in officially recorded hate-related incidents following implementation of Spain's Action Plan. Table 2 presents the full posterior inference results for the primary analysis including all post-intervention data (2019–2021).

During the post-intervention period, Spain recorded an average of 136.36 monthly hate crime incidents, compared to the model's counterfactual prediction of 113.34 incidents (SD = 5.73). This difference yields an absolute effect of 23.02 additional monthly incidents (95% CI: 11.86–34.33) and a cumulative effect of 828.86 additional

incidents over the 36-month post-intervention period (95% CI: 426.82–1,235.97). In relative terms, recorded hate crimes increased by 20.31% (95% CI: 10.46%–30.29%) compared to expected levels absent the intervention. The posterior probability of a policy-related effect is > 0.99 under the BSTS specification, indicating that the observed post-intervention divergence is unlikely to be explained by chance alone.

Table 2. Posterior Inference of Causal Impact including 2020 data: A comprehensive evaluation of the actual vs. predicted impact, including absolute and relative effects, along with their statistical significance.

Metric	Average	Cumulative
Observed	136.36	4909.0
Prediction (s.d.)	113.44 (5.97)	4080.14 (206.42)
95% CI	[102.03, 124.51]	[3673.91, 4516.52]
Absolute effect (s.d.)	23.02 (5.73)	828.86 (206.42)
95% CI	[11.86, 34.33]	[426.82, 1235.97]
Relative effect (s.d.)	20.31% (5.06%)	20.31% (5.06%)
95% CI	[10.46%, 30.29%]	[10.46%, 30.29%]

Note. Posterior probability of policy-related effect > 0.99. CI = credible interval.

To address potential confounding from COVID-19 pandemic effects-which may have altered both hate crime incidence and reporting behavior-we conducted a parallel analysis excluding 2020 data. As Table 3 shows, results remain robust and, notably, stronger. The observed average rises to 146.17 incidents against a prediction of 114.79 (SD = 6.72), yielding a relative effect of 27.34% (95% CI: 16.30%–39.24%). The posterior probability again exceeds 0.99. The larger effect size when excluding the pandemic year suggests that COVID-19 restrictions may have temporarily suppressed both hate crime incidents and reporting, attenuating the policy's observed impact during 2020.

Table 3. Posterior Inference of Causal Impact excluding 2020 data: Detailed evaluation highlighting the actual versus predicted impact, absolute and relative effects, and their statistical significance, omitting the pandemic's anomalous data

Metric	Average	Cumulative
Observed	146.17	3508.0
Prediction (s.d.)	114.79 (6.72)	2754.84 (161.26)
95% CI	[101.12, 127.46]	[2426.9, 3059.04]
Absolute effect (s.d.)	31.38 (6.72)	753.16 (161.26)
95% CI	[18.71, 45.05]	[448.96, 1081.1]
Relative effect (s.d.)	27.34% (5.85%)	27.34% (5.85%)
95% CI	[16.3%, 39.24%]	[16.3%, 39.24%]

Note. Posterior probability of policy-related effect > 0.99. Analysis excludes pandemic year to isolate policy effects from COVID-19 confounding.

Figures 1 and 2 visualize these findings. The top panels display observed versus predicted hate crime trends, with the shaded area representing the counterfactual prediction and its uncertainty. The divergence between observed (solid line) and predicted (dashed line) values becomes evident immediately following the intervention date. The middle panels show pointwise causal effects over time, while the bottom panels display cumulative effects, illustrating that the policy's impact accumulated progressively rather than appearing as a one-time shift.

Figure 1. Causal impact of Spain's Action Plan on recorded hate crimes (2015–2021). Top panel: Observed (solid line) versus counterfactual predicted (dashed line) monthly hate crime counts, with 95% credible interval (shaded). Middle panel: Pointwise causal effect estimates. Bottom panel: Cumulative causal effect over the post-intervention period.

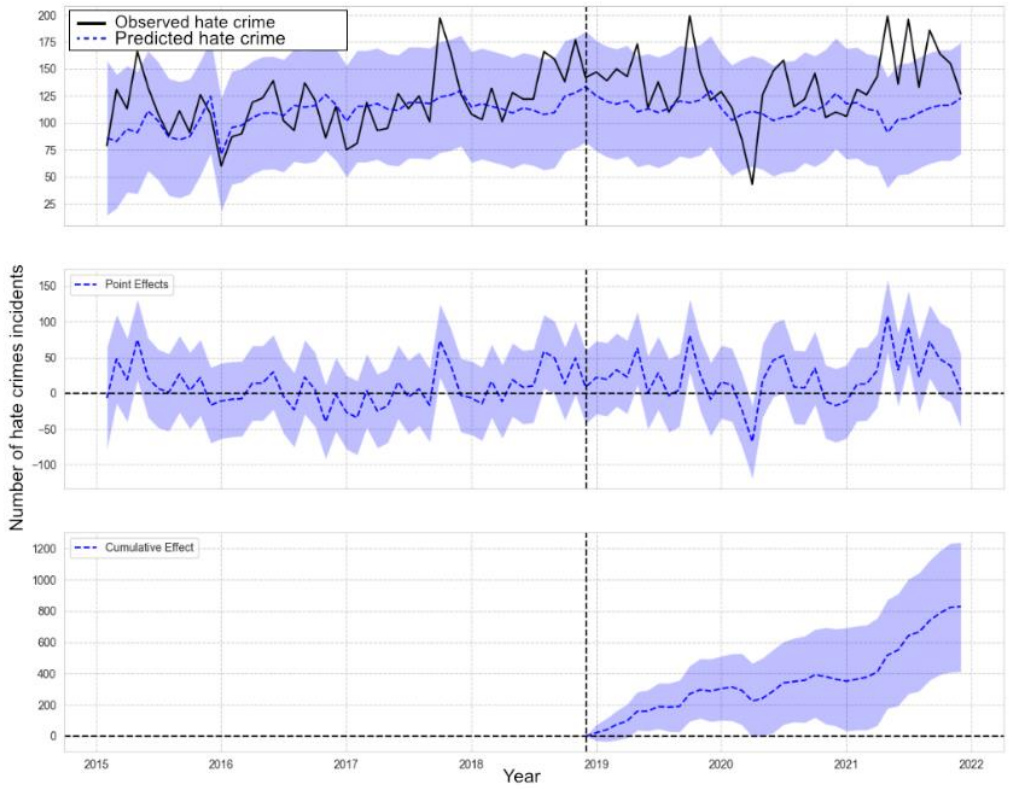
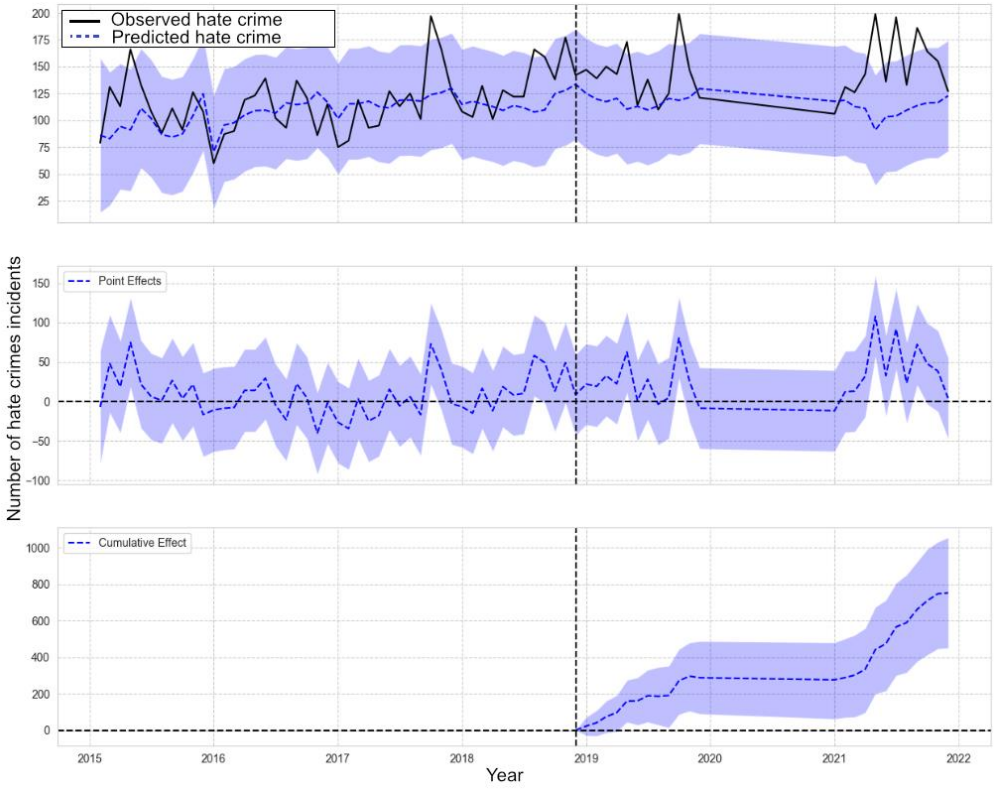


Figure 2. Causal impact analysis excluding 2020 pandemic year. Panels as in Figure 1. The larger effect size (27.34% vs. 20.31%) suggests COVID-19 restrictions attenuated the policy's observed impact during 2020.



Three descriptive features of the estimated post-intervention divergence merit attention. First, the gap between observed and predicted series appears early in the post-intervention period, which is compatible with a relatively rapid institutional response, although it does not by itself identify the mechanism through which the effect operated. Second, the divergence persists across the post-intervention window, suggesting that the estimated effect was not limited to a short-lived spike. Third, the cumulative gap increases over time, which is consistent with sustained changes in reporting, detection, recording, or some combination of these processes.

4.2 Robustness Checks

To validate the causal interpretation of our findings, we conducted the four robustness tests specified in Section 3.5. Table 4 summarizes the results.

Table 4. Refutation Test and Specificity Validation Results.

Test	Av. Abs. Effect	Std. Dev.	Post. Prob. of Causal Effect (%)
BSTS Analysis	22.41	6.02	100.0%
Placebo Treatment	-0.11	0.22	68.03%
Random Common Cause	22.75	5.62	100.0%
Subset Validation	22.28	7.03	99.9%

Note. Placebo test confirms no spurious effects. Random common cause and subset validation confirm robustness to omitted variables and sampling variability.

The primary BSTS analysis estimated an average absolute effect of 22.41 additional monthly incidents (SD = 6.02) with a posterior probability > 0.99. The placebo test, using a randomly generated outcome variable, yielded a negligible effect of −0.11 (SD = 0.22) with a posterior probability of only 68.03%-confirming that the model does not produce spurious causal effects when none exist. The random common cause test, which introduced a synthetic confounding covariate, produced results nearly identical to the primary analysis (22.75 ± 5.62, posterior probability > 0.99), demonstrating robustness to potential omitted variable bias. The subset validation, using 75% of

observations, yielded consistent estimates (22.28 ± 7.03, posterior probability 99.9%), supporting generalizability.

Table 5. Summary of Causal Impact Analysis for different slots for the Specificity Validation Test

Slot	Actual	Prediction	Abs. Effect	Rel. Effect	Post.Prob.	p
1	118.25	116.49 (7.93)	1.76 (7.93)	1.51% (6.81%)	59.24%	0.41
2	118.25	121.0 (7.71)	-2.75 (7.71)	-2.28% (6.37%)	63.64%	0.36

Note. Neither pre-intervention window shows significant effects, confirming that post-intervention findings are policy-specific.

Neither alternative time slot produced significant effects. Slot 1 yielded an absolute effect of 1.76 (SD = 7.93) with a posterior probability of 59.24% (p = 0.41), while Slot 2 showed an effect of −2.75 (SD = 7.71) with a posterior probability of 63.64% (p = 0.36). These null findings in placebo periods support the view that the post-intervention divergence is unlikely to reflect obvious pre-existing temporal patterns, seasonal structure, or model artefacts. Taken together, the validation tests are consistent with a policy-related effect on officially recorded hate-related incidents.

5. Discussion

5.1 Findings

This study provides evidence that Spain's Action Plan to Combat Hate Crimes (2019-2021) was associated with a sizeable increase in officially recorded hate-related incidents relative to the modelled counterfactual. Across specifications, the estimated increase ranges from approximately 20% to 27%, and the post-intervention divergence remains robust to the placebo and sensitivity checks reported above. Under the assumptions of the BSTS design, these findings are consistent with a policy-related effect on official recording rather than with random fluctuation alone.

The mechanism behind this increase cannot be established conclusively from the outcome variable itself. Our preferred interpretation is that the Action Plan improved reporting, detection, and recording rather than producing a genuine rise in hate-motivated victimisation. We favour this interpretation because the policy package explicitly targeted training, prevention, victim support, and institutional response; because both Spanish and EU-level evidence document substantial underreporting and recurrent barriers related to trust, awareness, and reporting pathways; and because the dependent variable used here captures official recording rather than direct victimisation prevalence (Ministry of Interior, 2019; FRA, 2021). This reading is reinforced when the findings are set against prior empirical work. The temporal signature estimated here—early onset and steady accumulation—differs from the pattern documented for genuine victimisation surges, which in the literature are event-triggered and follow spike-and-decay dynamics around focal events such as terrorist attacks or the Brexit referendum (Piatkowska & Lantz, 2021; Piatkowska & Stults, 2022); a sustained, trend-like divergence is instead what recording and reporting reforms would produce. The plausibility of the training channel is consistent with Johnson's (2023) finding that the clarity of written hate crime guidelines improves law enforcement identification of incidents—precisely what the 2020 update of the Spanish operational protocol standardised—while the victim-side channel is consistent with evidence that perceptions of police legitimacy and responsiveness shape reporting decisions (Wiedlitzka et al., 2018; FRA, 2021). More broadly, the result parallels the experience of England and Wales, where police-recorded hate crime more than tripled between 2013 and 2022 while victimisation-survey estimates declined, a divergence that the Home Office attributes primarily to improvements in police recording and identification rather than to rising victimisation (Home Office, 2024). Even so, alternative

explanations, such as concurrent cultural shifts, changing public sensitivity to bias motives, or changes in police classification practices, cannot be fully ruled out.

The stronger estimated effect observed when excluding the pandemic year should likewise be interpreted cautiously. COVID-19 plausibly altered opportunities for victimisation, willingness and ability to report, and institutional processing simultaneously. For that reason, we read the non-pandemic specification as a robustness exercise suggesting that the main finding is not driven by 2020 alone, rather than as definitive proof of the policy's "true" effect size.

This interpretation aligns with findings from the 2024 Spanish Victim and Witness Survey (Muniesa Tomás et al., 2024), which revealed that most hate crime victims still choose not to report incidents, citing mistrust in authorities and perceived institutional ineffectiveness. Notably, over 90% of survey respondents perceived an increase in hate incidents in recent years. Together, these findings suggest a complex reality: hate crimes may indeed be rising in absolute terms, but the recorded increase also reflects-perhaps primarily-enhanced societal awareness and improved institutional mechanisms for capturing incidents that previously remained invisible in official statistics.

5.2 Policy Implications

These findings suggest that awareness- and capacity-oriented interventions can improve the official visibility of hate-related incidents in settings marked by substantial underreporting. In the Spanish case, the Action Plan is compatible with increased reporting, detection, and recording within a system where a considerable share of victimisation likely remains outside official statistics. The main policy implication is therefore not simply that "more cases" were produced, but that institutional arrangements

affecting recognition, access, recording, and response deserve evaluation alongside strategies aimed at reducing offending itself (FRA, 2021; Ministry of Interior, 2019).

First, interventions that combine police training, victim support, and awareness-building may help reduce part of the dark figure of hate crime by improving the pathways through which incidents become visible to authorities. For policymakers, this means that the success of hate crime policy should not be assessed only in terms of whether underlying victimisation falls immediately, but also in terms of whether institutions become better able to identify, register, and respond to bias-motivated harm. In contexts where underreporting is extensive, improvements in official visibility can themselves constitute a meaningful policy achievement.

Second, the persistence of the estimated post-intervention divergence suggests that the effects of institutional interventions may be sustained rather than merely symbolic or short-lived. Unlike policies that produce a brief attention spike and then fade, the Action Plan appears compatible with a more durable strengthening of institutional response. This points to the importance of multi-year implementation horizons, sustained resourcing, and continued professionalisation, especially when policy objectives include not only prevention but also improved detection and support.

Third, policy evaluation in this field must distinguish carefully between incidence and visibility. An increase in officially recorded hate-related incidents after a policy intervention is not inherently either good or bad news; its meaning depends on the mechanism producing the change. If the increase reflects improved reporting, detection, and recording of previously hidden incidents, the intervention may be functioning as intended. If it reflects a genuine increase in underlying victimisation, more intensive primary prevention may be required. Our findings are more consistent with the former interpretation, but they also underline the need for complementary evaluation tools,

especially victimisation surveys and other data sources capable of distinguishing institutional visibility from underlying prevalence.

Fourth, international covariates can strengthen counterfactual construction when randomised evaluation is infeasible, but only if their role is stated carefully. In this study, Germany and Finland function as predictive covariates that help approximate Spain's unobserved post-2018 trajectory; they are not treated as fully harmonised legal equivalents of the Spanish framework. Comparative designs of this kind can be highly useful for future policy evaluations, provided their assumptions are stated as explicitly as their results.

Fifth, major external shocks such as COVID-19 can complicate the interpretation of policy effects and should be incorporated explicitly into evaluation design. The pandemic likely affected opportunities for victimisation, willingness and ability to report, and institutional processing at the same time. More broadly, this suggests that future evaluations of hate crime policy should routinely test the sensitivity of their conclusions to disruptive contextual events, rather than assuming that official trends can be interpreted independently of wider social interruptions.

5.3 Limitations

Several limitations warrant consideration when interpreting these findings.

A first limitation concerns measurement validity: our outcome measure-officially recorded hate-related incidents-captures only incidents that victims report and that authorities classify as hate-motivated. This measure cannot distinguish between changes in actual victimization, changes in reporting behavior, and changes in police classification practices. While we argue that improved reporting is the most plausible mechanism, we

cannot definitively rule out contributions from the other factors. Complementary victimization surveys conducted before and after policy implementation would strengthen causal claims.

Second, the temporal scope is limited: the post-intervention observation period (36 months, or 24 months excluding 2020) may be insufficient to capture longer-term dynamics. Policy effects could decay as initial attention fades, or they could strengthen as institutional changes become further embedded. Extended observation would clarify the durability of effects.

Third, Spain's specific institutional context, political climate, and baseline underreporting rate may limit the transferability of findings to other countries. The Action Plan's effectiveness likely depends on implementation fidelity, resource allocation, and pre-existing police-community relations-factors that vary across jurisdictions. Replication in other European contexts would strengthen external validity.

Finally, the design rests on the assumption, stated in Section 3.2, that the German and Finnish series would have tracked the Spanish outcome in a sufficiently informative way absent the intervention. Both countries were selected for the availability of continuous official monthly data and their participation in structured European monitoring frameworks, but they are predictive covariates rather than harmonised legal equivalents of Spain: differences in operational definitions and recording practices could introduce noise into the counterfactual, although they are unlikely to produce the systematic, Spain-specific, intervention-coincident bias needed to explain our findings. The robustness checks mitigate this concern but do not eliminate it.

6. Conclusion and Future Work

Spain’s Action Plan to Combat Hate Crimes (2019-2021) was associated with a substantial increase in officially recorded hate-related incidents relative to counterfactual

expectations. Under the assumptions of the BSTS design, this divergence is consistent with a policy-related effect that survived multiple robustness checks. The study therefore provides evidence that the Action Plan changed the visibility of hate-related incidents in official statistics, but it should not be read as a direct estimate of total underlying victimisation.

The most plausible substantive interpretation is that the Action Plan reduced part of the dark figure through improvements in reporting, detection, and recording. However, because the dependent variable is an institutional recording measure, the present design cannot fully disentangle reporting behaviour, police identification practices, classification decisions, and underlying incidence. The contribution of the study is therefore not to settle that question definitively, but to show that a national policy package focused on institutional response coincided with a marked and model-consistent change in official hate-related counts (Ministry of Interior, 2019; FRA, 2021).

For policymakers, the core lesson is clear: targeted interventions combining police training, community outreach, and victim support can meaningfully improve hate crime detection. However, increased recording is a means, not an end. The ultimate goal remains reducing victimization itself-a challenge that will require sustained commitment, continued monitoring, and complementary primary prevention strategies.

This study opens several avenues for further investigation. Most pressing is the need to disentangle reporting from incidence. Longitudinal victimization surveys, administered before and after policy implementation using consistent methodology, could directly measure whether the proportion of victims reporting incidents increased (supporting the awareness interpretation) or whether victimization rates themselves changed. Spain's 2024 Victim and Witness Survey represents a valuable starting point, but retrospective designs cannot fully address this question.

Comparative policy analysis across multiple EU countries implementing different intervention strategies could identify which specific components-police training, community outreach, victim support services, legal reforms-drive observed effects. Such analysis would require harmonized outcome measures and careful attention to implementation variation.

Finally, extended time-series analysis incorporating post-2021 data would clarify whether effects persist, decay, or evolve as the policy matures. The transition from active implementation (2019–2021) to routinized practice presents a natural opportunity to study sustainability.

Data availability statement

The analysis code and the derived Spanish monthly series used in the models are available at https://github.com/albertiny98/CausalImpactPublicPoliciesSpain. The monthly control series for Germany and Finland were obtained from third-party government authorities via official correspondence and cannot be publicly redistributed by the authors. Researchers may request access to these monthly series directly from the competent authorities; supporting correspondence details can be provided by the corresponding author upon reasonable request.

Acknowledgments

We extend our heartfelt gratitude to the Spanish Ministry of Interior, particularly to all members of the Statistical Analysis and Victim Support Area, for their invaluable support and cooperation, which have been pivotal to our research. Our thanks also go to the German Federal Ministry of Interior and the Finnish Ministry of Justice for their assistance and data provision. Special recognition is due to Carlos Fernández for his

unwavering dedication and expert insights that significantly contributed to the success of this study. Last but certainly not least, we express our profound appreciation to our spouses, Marta J. and Ana T., for their endless patience, understanding, and encouragement. Their support has been the bedrock of our perseverance and dedication throughout this endeavor.

References

Angrist, J. D., & Pischke, J.-S. (2009). Mostly harmless econometrics: An empiricist's companion. Princeton University Press. https://doi.org/10.2307/j.ctvcm4j72

Brodersen, K. H., Gallusser, F., Koehler, J., Remy, N., & Scott, S. L. (2015). Inferring causal impact using Bayesian structural time-series models. The Annals of Applied Statistics, 9(1), 247–274. https://doi.org/10.1214/14-AOAS788

Callaway, B., & Sant'Anna, P. H. C. (2021). Difference-in-differences with multiple time periods. Journal of Econometrics, 225(2), 200–230. https://doi.org/10.1016/j.jeconom.2020.12.001

Card, D., & Krueger, A. B. (2000). Minimum wages and employment: A case study of the fast-food industry in New Jersey and Pennsylvania. American Economic Review, 90(5), 1397–1420. https://doi.org/10.1257/aer.90.5.1397

Chai, K. (2025). A qualitative analysis of English universities' policies on addressing students' experiences of hate crime and incidents. *Policy Reviews in Higher Education, 9*(2), 300-323. https://doi.org/10.1080/23322969.2025.2474394

Clark-Moorman, K., Rydberg, J., & McGarrell, E. F. (2019). Impact evaluation of a parolee-based focused deterrence program on community-level violence. Criminal Justice Policy Review, 30(9), 1408–1430. https://doi.org/10.1177/0887403418812999

Dancygier, R. (2023). Hate crime supporters are found across age, gender, and income groups and are susceptible to violent political appeals. Proceedings of the National Academy of Sciences, 120(7), e2212757120. https://doi.org/10.1073/pnas.2212757120

Dorsett, R. (2021). A Bayesian structural time series analysis of the effect of basic income on crime: Evidence from the Alaska Permanent Fund. Journal of the Royal Statistical Society: Series A (Statistics in Society), 184(1), 179–200. https://doi.org/10.1111/rssa.12619

European Union Agency for Fundamental Rights. (2010). European Union Minorities and Discrimination Survey: Main results report. Publications Office of the European Union. https://doi.org/10.2811/88055

European Union Agency for Fundamental Rights. (2018). Hate crime recording and data collection practice across the EU. Publications Office of the European Union. https://fra.europa.eu/sites/default/files/fra_uploads/fra-2018-hate-crime-recording_en.pdf

European Union Agency for Fundamental Rights. (2021). Encouraging hate crime reporting: The role of law enforcement and other authorities. Publications Office of the European Union. https://doi.org/10.2811/303805

Farrell, A., & Lockwood, S. (2023). Addressing hate crime in the 21st century: Trends, threats, and opportunities for intervention. Annual Review of Criminology, 6, 107–130. https://doi.org/10.1146/annurev-criminol-030920-091908

Fiscalía General del Estado. (2019). Circular 7/2019, de 14 de mayo, sobre pautas para interpretar los delitos de odio tipificados en el artículo 510 del Código Penal. Boletín Oficial del Estado.

Herek, G. M. (2009). Hate crimes and stigma-related experiences among sexual minority adults in the United States: Prevalence estimates from a national probability sample. Journal of Interpersonal Violence, 24(1), 54–74. https://doi.org/10.1177/0886260508316477

Home Office. (2024). *Hate crime, England and Wales, year ending March 2024* (Statistical bulletin). https://www.gov.uk/government/statistics/hate-crime-england-and-wales-year-ending-march-2024

Imbens, G. W., & Rubin, D. B. (2015). Causal inference for statistics, social, and biomedical sciences: An introduction. Cambridge University Press. https://doi.org/10.1017/CBO9781139025751

Jacobs, J. B., & Potter, K. (1998). Hate crimes: Criminal law and identity politics. Oxford University Press.

Johnson, W. D. (2023). Optimizing written hate crime policy: An analysis of the impact content variation has on hate crime policy effectiveness (Doctoral dissertation, George Mason University). https://cls.gmu.edu/defenses/1430

Ley Orgánica 10/1995, de 23 de noviembre, del Código Penal. Boletín Oficial del Estado.

López Gutiérrez, J., Fernández Villazala, T., Máñez Cortinas, C. J., San Abelardo Anta, M. Y., & Gómez Esteban, J. (2021). *Informe de la encuesta sobre delitos de odio (Technical report)*. Ministerio del Interior, Gobierno de España.

Ministerio del Interior. (2019). Plan de Acción de lucha contra los Delitos de Odio. Gobierno de España.

Ministerio del Interior. (2020). *Protocolo de actuación de las Fuerzas y Cuerpos de Seguridad para los delitos de odio y conductas que vulneran las normas legales sobre discriminación* (Instrucción 4/2020, Secretaría de Estado de Seguridad; edición revisada, julio 2020). Gobierno de España.

Ministerio del Interior. (2022). *II Plan de Acción de lucha contra los delitos de odio*. Gobierno de España.

Ministerio del Interior. (2023). Sistema Estadístico de Criminalidad. Gobierno de España.

Ministerio del Interior. (2024). *Informe sobre la evolución de los delitos de odio en España 2024*. Gobierno de España.

Ministry of the Interior of Finland. (n.d.). Hate crime. Government of Finland.

Muniesa Tomás, M. P., Fernández Villazala, T., Máñez Cortinas, C. J., Colás Fuentes, E., Méndez Matos, G., Matilla Molina, A., González González, M., Amado Hernández, M. P., Gómez Esteban, J., & Buquerín Pascual, S. (2024). *Informe de la encuesta sobre delitos de odio: Víctimas y testigos*. Ministerio del Interior, Gobierno de España.

Organization for Security and Co-operation in Europe, Office for Democratic Institutions and Human Rights. (n.d.-a). National frameworks to address hate crime in Germany.

Organization for Security and Co-operation in Europe, Office for Democratic Institutions and Human Rights. (n.d.-b). National frameworks to address hate crime in Finland.

Pearl, J. (2009). Causality: Models, reasoning, and inference (2nd ed.). Cambridge University Press. https://doi.org/10.1017/CBO9780511803161

Piatkowska, S. J., Hövermann, A., & Yang, T. C. (2020). Immigration influx as a trigger for right-wing crime: A temporal analysis of hate crimes in Germany in the light of the "refugee crisis". The British Journal of Criminology, 60(3), 620–641. https://doi.org/10.1093/bjc/azz073

Piatkowska, S. J., & Lantz, B. (2021). Temporal clustering of hate crimes in the aftermath of the Brexit vote and terrorist attacks: A comparison of Scotland and England and Wales. The British Journal of Criminology, 61(3), 648–669. https://doi.org/10.1093/bjc/azaa090

Piatkowska, S. J., & Stults, B. J. (2022). Brexit, terrorist attacks, and hate crime: A longitudinal analysis. Social Problems, 69(4), 968–996. https://doi.org/10.1093/socpro/spab005

Vergani, M., Perry, B., Freilich, J., Chermak, S., Scrivens, R., Link, R., Kleinsman, D., Betts, J., & Iqbal, M. (2024). Mapping the scientific knowledge and approaches to defining and measuring hate crime, hate speech, and hate incidents: A systematic review. Campbell Systematic Reviews, 20(2), e1397. https://doi.org/10.1002/cl2.1397

Wiedlitzka, S., Mazerolle, L., Fay-Ramirez, S., & Miles-Johnson, T. (2018). Perceptions of police legitimacy and citizen decisions to report hate crime incidents in Australia. International Journal for Crime, Justice and Social Democracy, 7(2), 91–106. https://doi.org/10.5204/ijcjsd.v7i2.483

Williams, M. L., Burnap, P., Javed, A., Liu, H., & Ozalp, S. (2020). Hate in the machine: Anti-Black and anti-Muslim social media posts as predictors of offline racially and religiously aggravated crime. *The British Journal of Criminology, 60*(1), 93-117. https://doi.org/10.1093/bjc/azz049