AUTHORS DETAILS

Laura Nieto Torrejón – lnieto@ucam.edu

Universidad Católica San Antonio de Murcia

ORCID: 0000-0002-6206-585X

Nicolás Gonzálvez Gallego – ngonzalvez@ucam.edu

Universidad Católica San Antonio de Murcia

ORCID: 0000-0002-0230-1195

María Concepción Pérez Cárceles - concepcion.perez@cud.upct.es

Universidad Politécnica de Cartagena – Centros Universitarios de la Defensa.

ORCID: 0000-0002-7262-3413

Do search queries predict violence against women? A forecasting model based on Google trends.

1. Introduction

The elimination of gender-based violence (GBV) is a main concern in Europe and, consequently, in Spain. In 2004, the National Parliament passed the Act on Integrated Protection Measures Against Gender-based Violence, approved by all the parliamentary groups. According to the Convention in Instabul (Council of Europe, 2011), there are different forms of GBV, such as intimate partner violence (IPV), sexual harassment, sexual violence, female genital mutilation, stalking and forced marriage, among others. This paper, as well as the Spanish Act on GBV, deals with on IPV against women. Also, IPV is the type of GBV most extensively covered by European researchers (Bradbury-Jones et al., 2017).

According to the last national evidences, between 8.1% and 25.4% of Spanish women have suffered any kind of GBV at some time of their lives (Government of Spain, 2015), while other international organizations estimated that the percentage is around 10% (OECD, 2019). In terms of femicides, from 2003 to 2019, more than 1,000 women in Spain were killed by their partners or ex-partners (Government of Spain, 2019). This is not only a Spanish issue, but also a global priority. International studies revealed 1 in 3 women worldwide suffered intimate partnership violence and 38% of femicides are related to IPV episodes (World Health Organization, 2013). Indeed, the 2030 Agenda for Sustainable Development sets out the elimination of all types of violence against women, including GBV, as one of its main goals (United Nations, 2015).

Research on forecasting and GBV is mainly focused on predicting risk for an individual woman of suffering a violent incident. Several tools have been designed and implanted, such as Danger Assessment, B-Safer, ODARA, DVSI or SARA (Campbell et al., 2009; Kropp et al., 2005; Hilton et al., 2004; Williams and Houghton, 2004; Kropp and Hart, 2000). From a criminal research perspective, forecasting is receiving increasing attention. Predictive policing is an approach based on crime reduction to predict where crime is likely to occur and prevent its occurrence (Bretton, 2011; Drawve, 2014), which is leading to a shift from reactive to predictive policies (Pearsall, 2010). Police and researchers are implementing several hot spot mapping techniques as well as the Risk Terrain Modelling (RTM) approach in order to predict crime locations (Drawve, 2014). Although , as far as we know, there are no applications of

these techniques for IPV or GBV, they have been applied for robberies (Flaxman, 2014) or organized crime homicides (Dugato and Calderoni, 2017).

Those techniques have in common the purpose of predicting crime rates in the short term in small and very restricted geographical areas. Our aim, although frameworked within the forecast of IPV against women, is to predict it as a macro-phenomem for a whole country. We compare predictive capacity of different varialbes: fatalities, calls to national IPV helpline and Google queries. While the two former are more traditional measures, the latter is a more innovative indicator. Then, this study aims to contribute towards the use of Google and other user data to better understand IPV, so that public institutions can react in advance and reduce its effect. This set of predictors will allow them to identify which one has the most powerful predictive power for different forecast horizons. As stated by Stephens-Davidowitz (2013), the use of Google searches to proxy domestic violence and crime seems promising.

Google search queries were first used by Ginsberg (2009). The author found that influenza-related queries surveyed influenza activity in large population in real-time. Since then, researchers have tested the predictive power of Google data in different fields, such as macroeconomic series (Niesert et al., 2019; Choi and Varian, 2012), unemployment levels (D'Amuri and Marcucci, 2017), housing prices and sales trends (Wu and Brynjolfsson, 2017) or daily gains of crypto-currencies (Hotz-Behofsits, 2018).

Apart from forecasting economic variables, Google data have been extended to other research areas, some of them related to crime. Stephens-Davidowitz (2013) considered Google queries to estimate the victims of child abuse during the Great Recession in the United States. Kostakos (2018) analized public perceptions on terrorism, mafia and organized crime by using Google data. More recently, Gamma et al., (2019) studied how Google data can be helpful to predict methamphetamine-related crime in Germany, Switzerland and Austria. They found high cross-correlations between search volumes and crime statistics, although they did not get uniform results for all the countries and other variables.

2. Data

2.1. IPV data

Data of reported IPV in Spain has been collected from 2009 to 2018 on a quarterly basis, so that all the available information from the General Council of the Judiciary has been considered for this study. Reported IPV is split in two groups: with

an attached police statement (IPV_1) and without it (IPV_2). The former is made up of those IPV cases reported immediately after a violent or risky incident occurs, while the latter includes cases reported by the victim or third parties but not directly linked to a specific violent episode. This dual analysis with two samples of different but deeply related cases will lead to a more consistent quantitative analysis and, consequently, to more robust results. Ultimately, this will reinforce the validity of our Google-based index as a predictor of IPV.

2.2. Google Trends data

The exogenous variables of this study are Google Index for IPV (GI-IPV), fatalities and calls to national IPV helpline. GI-IPV, our first dependent variable is an indicator based on searches related to IPV through Google search engine from users located in Spain. In this country, from January 2009 to December 2018, Google's average market share was 96.18%, according to Statcounter's website. Then, Google widely represents the Spanish search engine market.

The analyzed search queries were selected in two steps. First, we entered a keyword related to the IPV helpline at Ubersugest, a free SEO tool, to look for alternative keywords. Second, among them, we chose the two most searched keywords more likely to be related with looking for information to cope with IPV (i.e. we considered that a user who is suffering IPV, or is aware of somebody in this situation, is more likely to search for "IPV helpline" than for "IPV statistics").

Then, Google Trends was used to get data on the relative popularity of the three keywords aggregated as a single search query group. Google Trends, available since 2004, provides monthly time series of search activity for a search term (or group of search terms, as in this paper) in a certain geographic area for a defined period of time. The index provided by Google Trends varies from 0 to 100, where 100 represents the specific month in which the search activity for the keyword(s) was the highest during the analyzed period (i.e. a value of 60 in a given month means that search volume was 60% of the highest value within the time series). To know more about how Google Trends works, see D'Amuri and Marcucci (2017) and Stephens-Davidowitz and Varian (2014).

Fatalities, our second dependent variable, are defined as the number of mortal victims derived from IPV. Finally, the third variable, calls to national IPV helpline is the number of Spanish Ministry of presidency and equality. Monthly data has been aggregated on a quarterly basis.

3. Models and methodology

We study different linear regression models to forecast quarterly reported IPV in Spain. First, we estimate a simple autoregressive model where the lagged quarterly reported IPV is the independent variable. This is our baseline model:

$$y_{t+h} = \beta_0 + \beta_1 y_t + \eta_{t+h}, \qquad t = 1,2,\dots,T \qquad (1)$$

Where y denotes the quarterly volume of IPV cases and h is the forecast horizon. Then, we estimate the following standard forecast model (Stock and Watson, 2003):

$$y_{t+h} = \beta_0 + \beta_1 y_t + \beta_2 x_t + +\varepsilon_{t+h}, \qquad t = 1,2,\dots,T \qquad (2)$$

Where a new variable x is added in order to analyze if the predictive power of these indicators improve the baseline model. Fatalities, GI-IPV and calls to national IPV helpline are tested as explicative variables.

In order to determine to what extent the predictor variables improve the baseline model, we calculate the relative reduction in the unexplained variance through the incremental R^2. In addition, a t-statistic tests whether the coefficients of the respective indicators are null. Consequently, the statistical significance of the relative reduction in unexplained variance is measured. Then, each of the predictive models are assessed by their respective ratio of the root mean squared forecast errors (RMSFE) over the baseline model´s one. The significance of the forecasting accuracy is determined by the Diebold-Mariano (1995) (MDM) and the Harvey-Leybourne-Newbold (1997) (HLN) test statistics.

The MDM test proposes an approach for testing equal forecast accuracies that is valid for potentially contemporaneously correlated, serially correlated and normal forecast errors. It is based on testing a zero mean in a series defined as the difference between the two forecasts' error loss function, which is named "loss differential" (d_t). The DM statistic under null hypothesis is:

$$DM = \frac{\bar{d}}{\sqrt{\hat{v}\bar{d}}} \rightarrow N(0,1)$$

$$\bar{d} = \frac{\sum_{t=1}^{h} d_t}{h}$$

$$\hat{v}\bar{d} = \frac{\widehat{Y_0} + 2\sum_{k=1}^{h-1} \widehat{Y_k}}{h}$$

$$\widehat{Y_k} = \frac{\sum_{t=k+1}^{h} (d_t - \bar{d})(d_{t-k} - \bar{d})}{h}$$

Where h is the forecast horizon length, $\hat{Y}_0$ is the estimated variance of d_t and $\hat{Y}_k$ is the estimated $k-th$ auto-covariance of d_t.

The tend to fail null hypothesis in small samples of the MDM is overcame by Harvey et al., (1997). They propose a forecast-encompassing test based on a DM-type approach, where the loss differential is redefined to permit the testing of an encompassing null hypothesis. This statistic follows a Student's t-distribution with n-1 degrees of freedom and shows whether differences in RMSFEs are statistically significant in augmented models.

$$HNL = \frac{\bar{d}}{\sqrt{Y_0+2Y_1+\cdots+2Y_{h-1}}} \cdot \left(n+1-dj+\frac{j(j-1)}{h}\right)$$

4. Results

Table 1 shows the descriptive statistics for the dependent variables and the indicators added to the baseline model. Regarding the dependent variables, the average quarterly reported cases is higher for IPV_2 than for IPV_1. As variation coefficient supports, mean values are representative (0.14 and 0.13 respectively). Concerning the regressors, fatalities and GI-IPV, their variation coefficients are higher (0.30 and 0.44, respectively) than calls to national IPV (0.16). This implies that the mean value is only representative in the last case.

Table 1. Descriptive statistics

	Mean	Std. Dev.	Var. Coeff.	Min.	Max.
Dependent variables					
IPV_1	27,883.69	3,955.77	0.14	22,504.00	36,129.00
IPV_2	31,008.26	3,880.97	0.13	26,166.00	39,415.00
IPV Indicators					
Fatality IPV	14.15	4.23	0.30	7.00	24.00
GI – IPV	33.18	14.65	0.44	13.00	77.67
Helpline	17,719.59	2,822.90	0.16	12,149.00	25,178.00

Table 2. Forecasting models estimation

	Baseline Model		GI Model			Fatality based Model			Helpline Model		
					IPV_1						
	β_1	R^2	β_1	β_2	R^2	β_1	β_2	R^2	β_1	β_2	R^2
h=1	0.930***	0.780	0.793***	61.660**	0.803	0.926***	-52.220	0.781	0.915***	0.036	0.779
h=2	0.880***	0.601	0.598***	118.280***	0.678	0.873***	-29.877	0.590	0.735***	0.334	0.631
h=3	0.971***	0.652	0.667***	130.320***	0.749	0.965***	-32.827	0.643	0.727***	0.518***	0.739
h=4	1.060***	0.739	0.934***	57.990	0.752	1.043***	-137.559	0.750	0.844***	0.425***	0.796
					IPV_2						
	β_1	R^2	β_1	β_2	R^2	β_1	β_2	R^2	β_1	β_2	R^2
h=1	0.919***	0.766	0.779***	63.010**	0.787	0.915***	-55.335	0.763	0.913***	0.018	0.759
h=2	0.869***	0.578	0.589***	116.730***	0.658	0.866***	-13.212	0.566	0.759***	0.293*	0.603
h=3	0.957***	0.611	0.639***	134.570***	0.722	0.952***	-24.931	0.600	0.748***	0.504***	0.707
h=4	1.070***	0.710	0.921***	67.510*	0.732	1.052***	-12.375	0.718	0.854***	0.473***	0.796

Significance levels: *0.1, **0.05 and ***0.01

Table 2 reports estimated coefficients and goodness of fit of the indicators for the proposed models (adjusted R^2). The baseline model shows high significance levels for all the coefficients regardless of the forecast horizon. Variance percentage explained by this model is, at least, 0.601 and 0.578 for IPV_1 and IPV_2, respectively. Then, we assess if alternative models' estimation and goodness of fit improve compared to the baseline model. According to R^2 values, GI-IPV model fits better than the baseline model for all the forecast horizons. The R^2 average increase for IPV_1 is 5.3% while for IPV_2 is 5.9%. Most relevant improvements are observed for the third forecast horizon (9.7% and 11.1% respectively). In addition, regression coefficients of the GI-IPV model are highly significant, with the single exception of the more distant forecast horizon for IPV_1.

In regards to fatality model, the R^2 increases in the first and fourth forecast horizon in IPV_1 and in the forecast horizon 4 in IPV_2. Nevertheless, the added regression coefficients are not significant for any of the horizons.

Helpline model shows a better fit than the baseline model for all the forecast horizons except for the first one in both cases (IPV_1 and IPV_2). The R^2 average increase for IPV_1 is 4.32% while for IPV_2 is 5.0%. The most relevant improvements are observed for the fourth forecast horizon (7.71% and 12.11%, respectively). Regression coefficients are significant in forecast horizons 3 and 4 for IPV_1 and IPV_2. Additionally, for IPV_2, the coefficient is also significant in the second horizon. The helpline model in the fourth forecast horizon is the one that shows the best adjustment.

Table 3 shows the results of the assessment for the augmented models relative to the baseline model. We calculate the relative RMSFE (Rel. RMSFE) obtained for the three models. A value below one means that the augmented model has a lower RMSFE than the baseline one. Information of MDM and HLN tests are also included in this table.

Table 3. Relative performance (baseline model versus augmented models)

	GI model			Fatality based model			Helpline model		
	Rel. RMSFE	MDM	HNL	Rel. RMSFE	MDM	HNL	Rel. RMSFE	MDM	HNL
				IPV_1					
h=1	0.943	-1.612	-1.712*	0.993	-0.498	-0.529	0.998	-0.548	-0.583
h=2	0.885	-1.985**	-2.111**	0.999	-0.142	-0.151	0.948	-1.294	-1.377
h=3	0.836	-2.800***	-3.045***	0.998	-0.247	-0.268	0.854	-1.719*	-1.869*
h=4	0.488	-2.318**	-2.475**	0.694	-1.451	-1.548	0.283	-3.548***	-3.786***
				IPV_2					
h=1	0.941	-1.279	-1.359	0.992	-0.549	-0.584	0.999	-0.511	-0.542
h=2	0.887	-1.708*	-1.817*	0.999	-0.070	-0.074	0.956	-1.174	-1.248
h=3	0.833	-2.986***	-3.181***	0.999	-0.309	-0.329	0.854	-1.698*	-1.809*
h=4	0.947	-1.685*	-1.799*	1.211	0.925	0.987	0.868	-1.663*	-1.775*

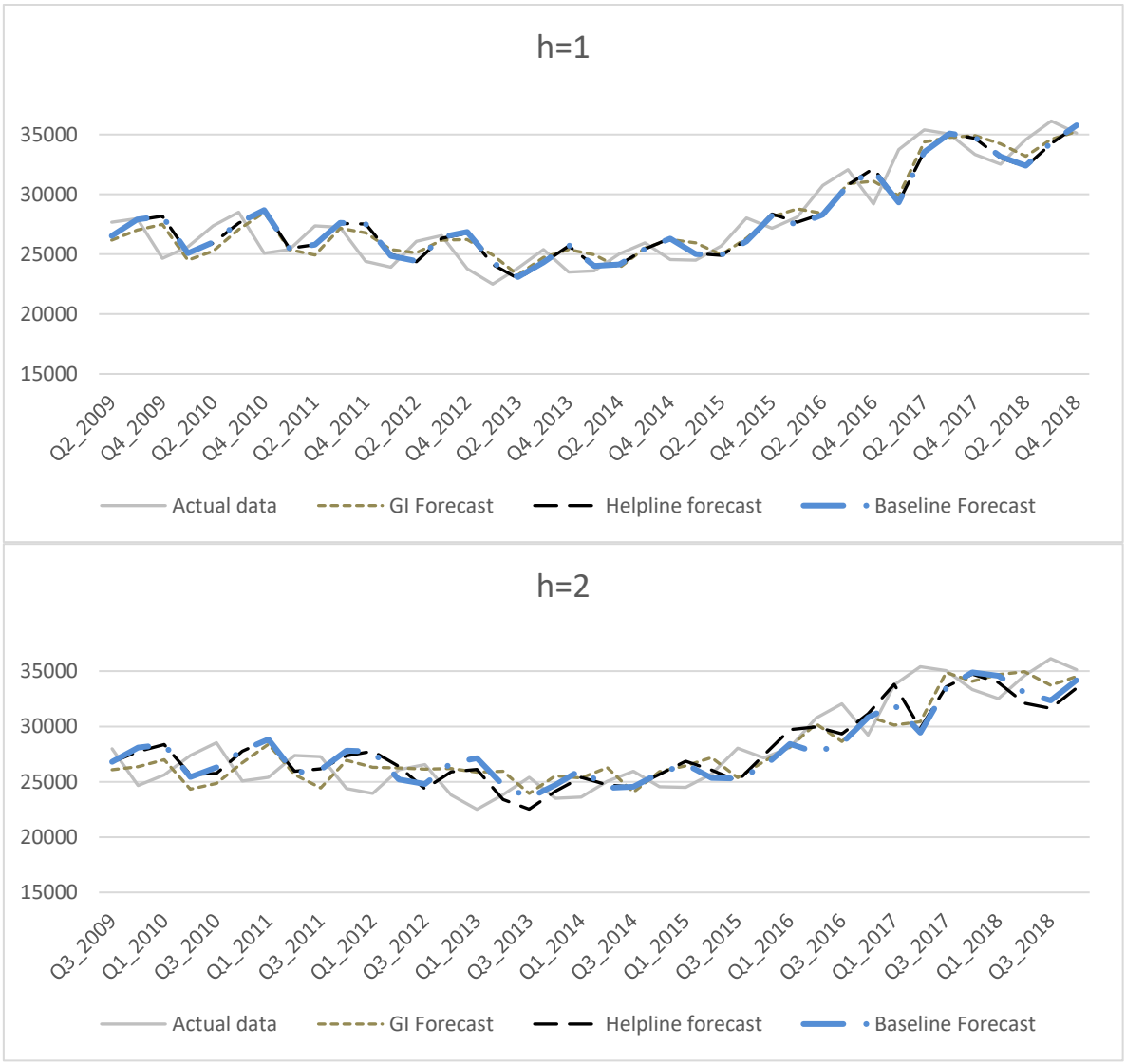
Significance levels: *0.1, **0.05 and ***0.01

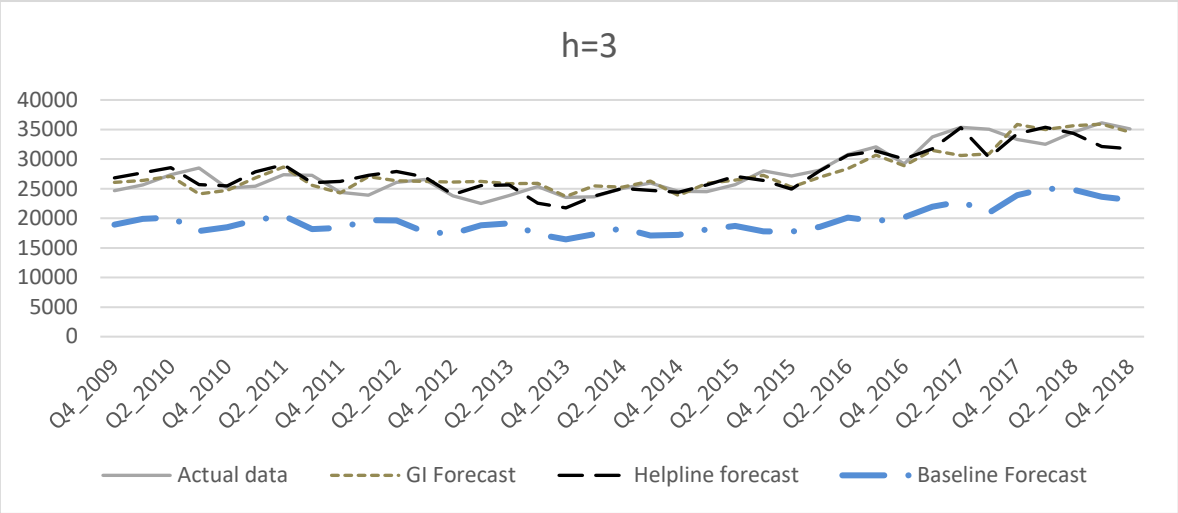
For IPV_1, relative RMSFE is below 1 for all the forecasting horizons. However, distances from relative RMSFE values to 1 are notably higher for GI-IPV model than for the basic model for forecast horizons from 1 to 3. For the forecast horizon 4 the greatest distance is presented by helpline model, especially for IPV_1. In line with these results, the improvement of the predictive power is statistically significant for all the periods in the case of GI-IPV model. For forecast horizon 4, the highest and most significant increase is shown by the helpline model, which beats the other three. These results denote that predictions based on Google data are more reliable to make 3 to 9 months forecasts while the helpline model is more reliable in the long-run.

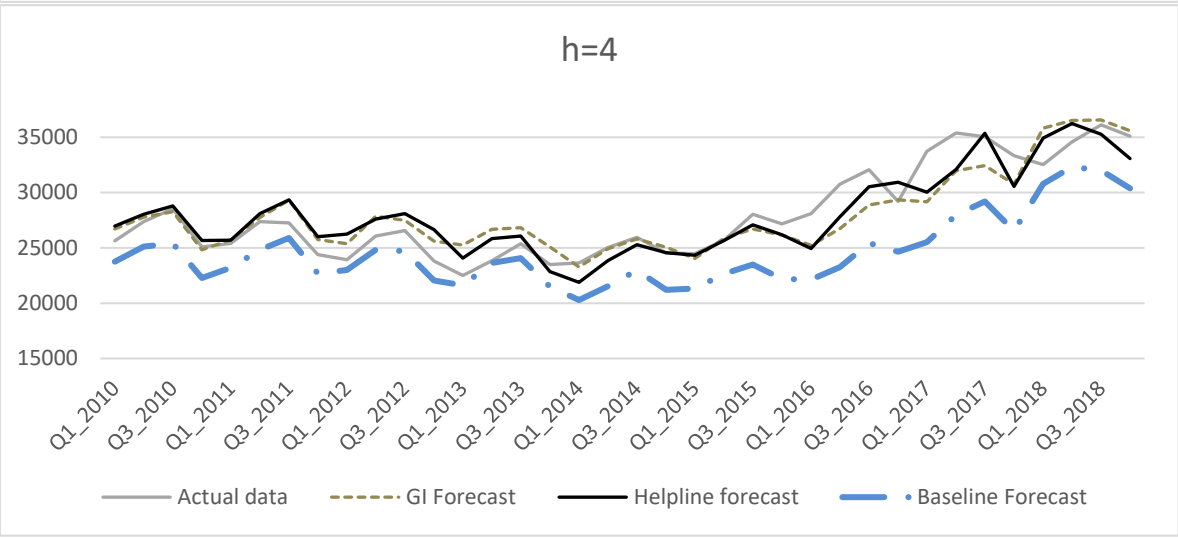
Regardless of the type of reported IPV, the change is not statistically significant for fatality-based model.

In Figure 1 the actual data series of reported IPV with police statement attached (IPV_1) is graphically confronted to the forecasted values obtained by the baseline, GI-IPV and helpline prediction models for the four forecast horizons. In line with statistical indicators offered in Table 3, we can observe how GI-IPV model progressively increases its forecasting accuracy throughout the analyzed period.

Figure 1: Actual data and forecasted values of reported IPV_1

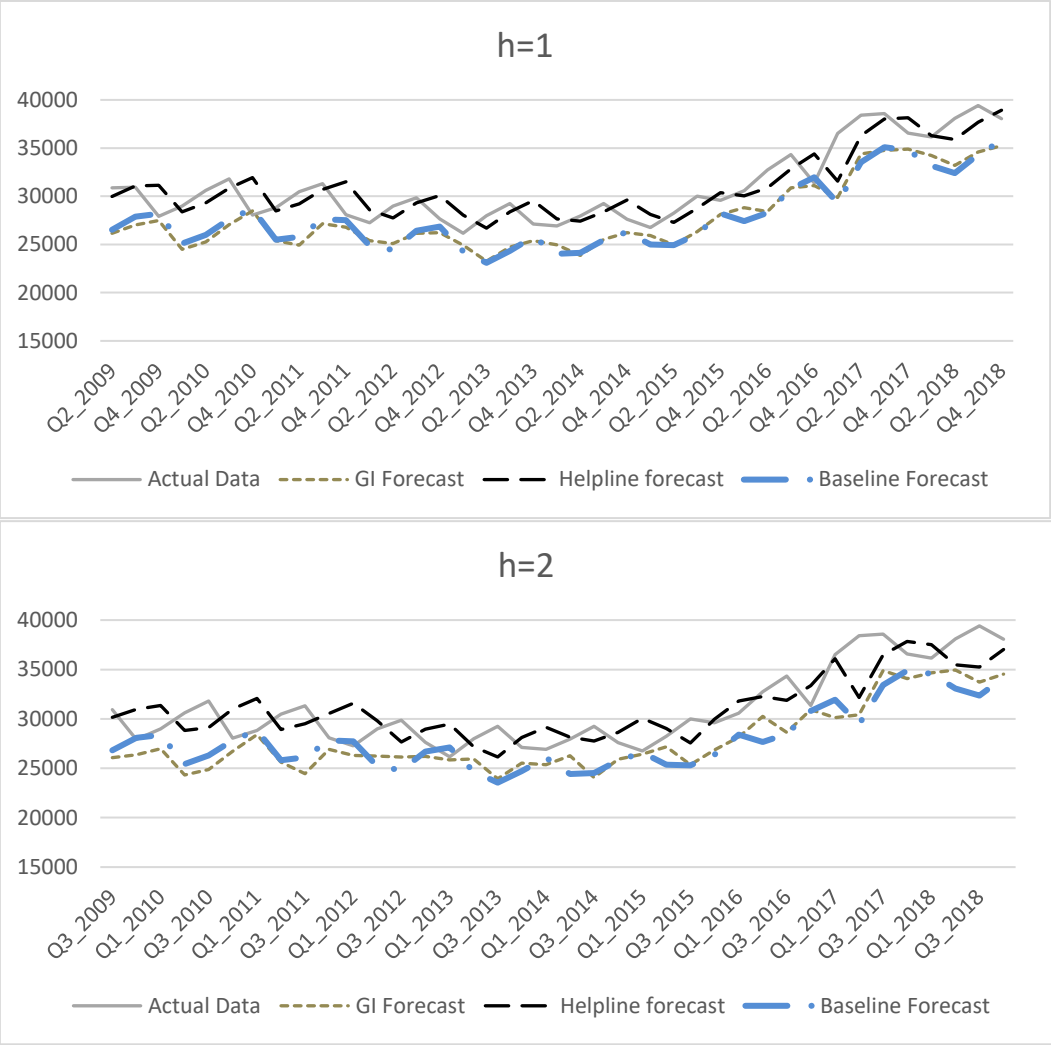


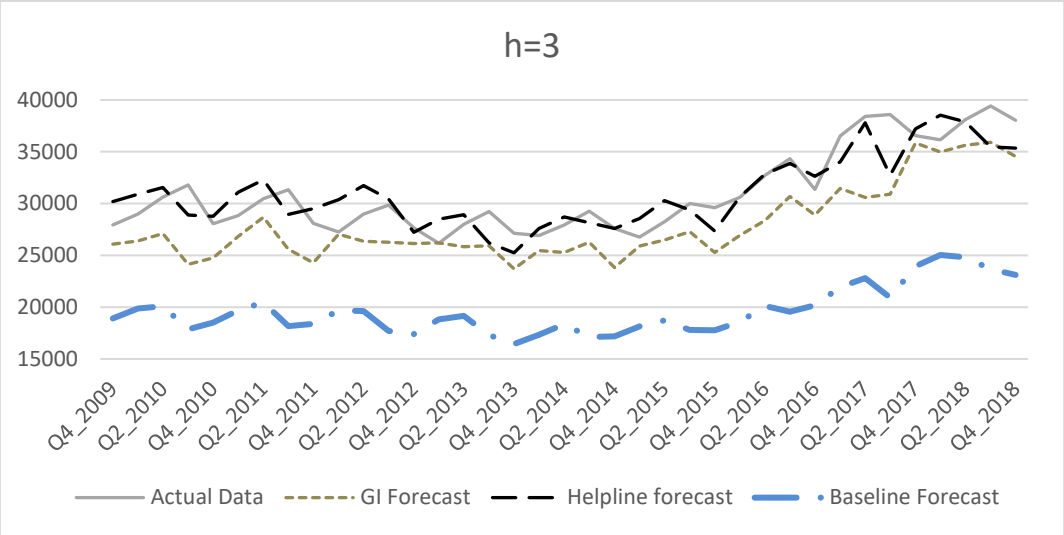


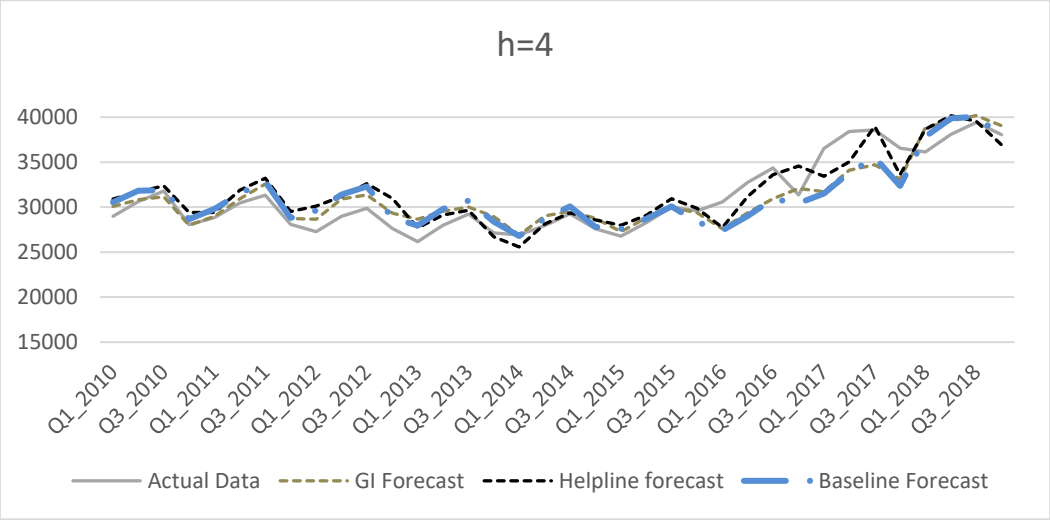


Similarly, Figure 2 shows actual values for IPV reported without a police statement together with forecasted values from the baseline, GI-IPV and Helpline models for all the forecast horizons.

Figure 2: Actual data and forecasted values of reported IPV_2







In-sample forecast

After running the out-sample analysis, as a robustness check, we also conduct an in-sample forecasting experiment using quarterly data from 2009 to 2017. It has been performed exclusively with GI-IPV model since this is the model that shows the highest forecast accuracy compared to the baseline model. Following Reinhard and Timmermann (2012) we have conducted the in-sample forecast at 10% and 90% split points of the sample. Table 4 shows the level of adjustment of GI-IPV model:

Table 4: In sample forecast for GI-IPV model. Goodness of fit (R^2).

	h=1	h=2	h=3	h=4
IPV_1	0.716	0.533	0.642	0.729
IPV_2	0.671	0.470	0.564	0.687

The goodness of fit of IPV_1 ranges from 0.533 (h=2) to 0.729 (h=4) and, for IPV_2, from 0.470 (h=2) and 0.687 (h=4). These values confirm the the predictive power of GI-IPV model over the time.

5. Conclusions

In this paper, we studied how different models predict reported intimate partner violence against women in Spain for the period 2009-2018 on a quarterly basis. We split reported cases in two groups: those immediately reported after a violent incident and those not directly connected to a specific episode of violence. Each of the three proposed models, tested to predict both types of cases, are based on a different indicator: fatalities, calls to the national IPV helpline and Google queries related to IPV. In order to test models' forecast accuracy, we run a lineal regression per indicator considering four forecast horizons, from one to four quarters and then, we compared it to a baseline model that solely includes the lagged dependent variable. Finally, we studied if the improvement in the predictive capacity of one model is statistically significant to determine which indicator makes the most accurate forecast.

Our results reveal that, regardless of the way IPV is reported, GI-based model beats all the other models for all the forecast horizons, with the single exception of h=4 (four quarters). On the other hand, the number of calls received by the national helpline is the best indicator to predict future reported IPV cases in the long term, since it is the best indicator for the most distant forecast horizon. Finally, it is relevant to mention that fatalities were not found to be a valid predictor for any of the forecast periods.

Those findings lead to different implications for public agencies in terms of the design, management and implementation of IPV policies. First, public institutions can use GI to estimate the responsiveness that will be required in the short term, especially in the next three months, since GI is the best indicator to forecast IPV cases three months away. Also, they can use GI to adjust the efforts in public campaigns that disclose information on support services for IPV's victims services. For this purpose, but in the long term (1 year), public officers should use the number of calls received by the national helpline. Second, GI has the highest predictive power for reported IPV cases with attached police statement because of a violent incident. It leads us to suggest that adding a Google-related item to current risk assessment tool could improve risk estimation.

To sum up, our findings are relevant to improve academics and public agencies' ability to forecast the volume of IPV against women at national level. However, this exploratory study must be replicated and confirmed in different countries. At the same

time, we encourage other researchers to consider different measures disclosed on a monthly basis in order to get more accurate forecasts that may have a more immediate impact on short term decisions.