Supplementary information for "Posterior calibration and exploratory analysis for natural language processing models" (EMNLP 2015)

Khanh Nguyen and Brendan O'Connor

August 2015

The original paper and this document are available at:
http://brenocon.com/nlpcalib

1 Sampling a deterministic function of a random variable

In several places in this paper, we define probability distributions over deterministic functions of a random variable, and sample from them by applying the deterministic function to samples of the random variable. This should be valid by construction, but we supply the following argument for further justification.

X is a random variable and $g(x)$ is a deterministic function which takes a value of X as its input. Since g depends on a random variable, $g(X)$ is a random variable as well. The distribution for $g(X)$, or aspects of it (such as a PMF or independent samples from it) can be calculated by marginalizing out X with a Monte Carlo approximation. Assuming g has discrete outputs (as is the case for the event counting function n, or connected components function CC), we examine the probability mass function:

$$\text{pmf}(h) \equiv P(g(X) = h) = \sum_x P(g(x) = h \mid x)\, P(x) \tag{1}$$

$$= \sum_x 1\{g(x) = h\} P(x) \tag{2}$$

$$\approx \frac{1}{S} \sum_{x \sim P(X)} 1\{g(x) = h\} \tag{3}$$

Eq. 2 holds because $g(x)$ is a deterministic function, and Eq. 3 is a Monte Carlo approximation that uses S samples from $P(x)$.

This implies that a set of g values calculated on x samples, $\{g(x^{(s)}) : x^{(s)} \sim P(x)\}$, should constitute a sample from the distribution $P(g(X))$; in our event analysis section we usually call this the "posterior" distribution of $g(X)$ (the $n(t, c)$

function there). In our setting, we do not directly use the PMF calculation above; instead, we construct normal approximations to the probability distribution $g(X)$.

We use this technique in several places. For the calibration error confidence interval, the calibration error is a deterministic function of the uncertain empirical label frequencies p_i; there, we propagate posterior uncertainty from a normal approximation to the Bernoulli parameter's posterior (the p_i distribution under the central limit theorem) through simulation. In the coreference model, the connected components function is a deterministic function of the antecedent vector; thus repeatedly calculating $\mathbf{e}^{(s)} := CC(\mathbf{a}^{(s)})$ yields samples of entity clusterings from their posterior. For the event analysis, the counting function $n(t, c, \mathbf{e}_{d(t)})$ is a function of the entity samples, and thus can be recalculated on each—this is a multiple step deterministic pipeline, which postprocesses simulated random variables.

As in other Monte Carlo-based inference techniques (as applied to both Bayesian and frequentist (e.g. bootstrapping) inference), the mean and standard deviation of samples drawn from the distribution constitute the mean and standard deviation of the desired posterior distribution, subject to Monte Carlo error due to the finite number of samples, which by the central limit theorem shrinks at a rate of $1/\sqrt{S}$. The Monte Carlo standard error for estimating the mean is $\sigma/\sqrt{S}$ where σ is the standard deviation. So with 100 samples, the Monte Carlo standard error for the mean is $\sqrt{100} = 10$ times smaller than standard deviation. Thus in the time series graphs, which are based on $S = 100$ samples, the posterior mean (dark line) has Monte Carlo uncertainty that is 10 times smaller than the vertical gray area (95% CI) around it.

2 Normalization in the coreference model

Durrett and Klein (2013) present their model as a globally normalized, but fully factorized, CRF:

$$P(\mathbf{a}|x) = \frac{1}{Z} \prod_i \exp(\mathbf{w}^\mathsf{T} \mathbf{f}(i, a_i, x))$$

Since the factor function decomposes independently for each random variable a_i, their probabilities are actually independent, and can be rewritten with local normalization,

$$P(\mathbf{a}|x) = \prod_i \frac{1}{Z_i} \exp(\mathbf{w}^\mathsf{T} \mathbf{f}(i, a_i, x))$$

This interpretation justifies the use of independent sampling to draw samples of the joint posterior.

3 Event analysis: Corpus selection, country affiliation, and parsing

Articles are filtered to yield a dataset about world news. In the New York Times Annotated Corpus, every article is tagged with a large set of labels. We include articles that contain a category whose label starts with the string *Top/News/World*, and exclude articles with any category matching the regex */(Sports|Opinion)*, and whose text body contains a mention of at least one country name.

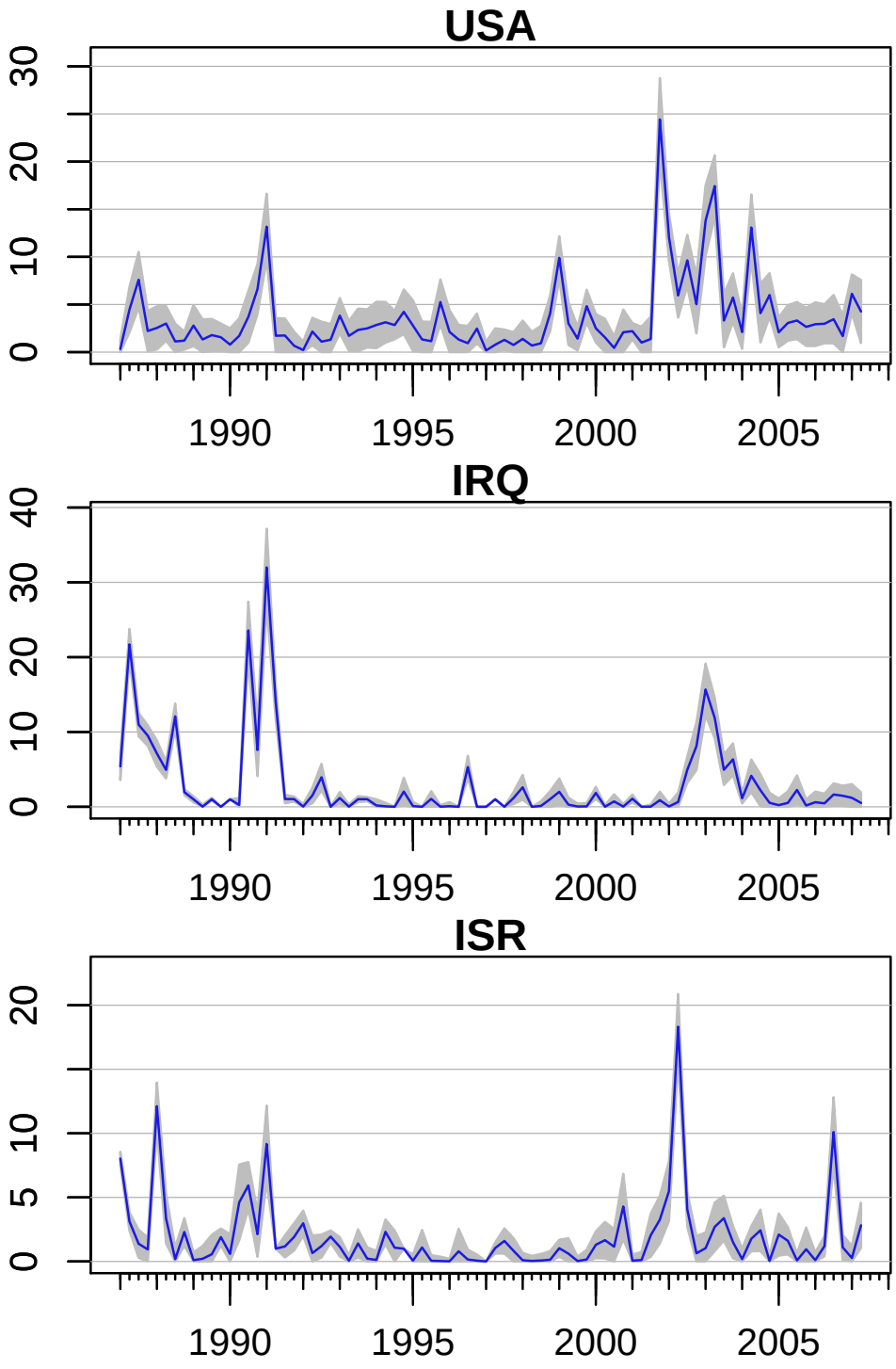
Country names are taken from the dictionary *country_igos.txt* based on previous work (http://brenocon.com/irevents/). Country name matching is case insensitive and uses light stemming: when trying to match a word against the lexicon, if a match is not found, it backs off to stripping the last and last two characters. (This is usually unnecessary since the dictionary contains modifier forms.)

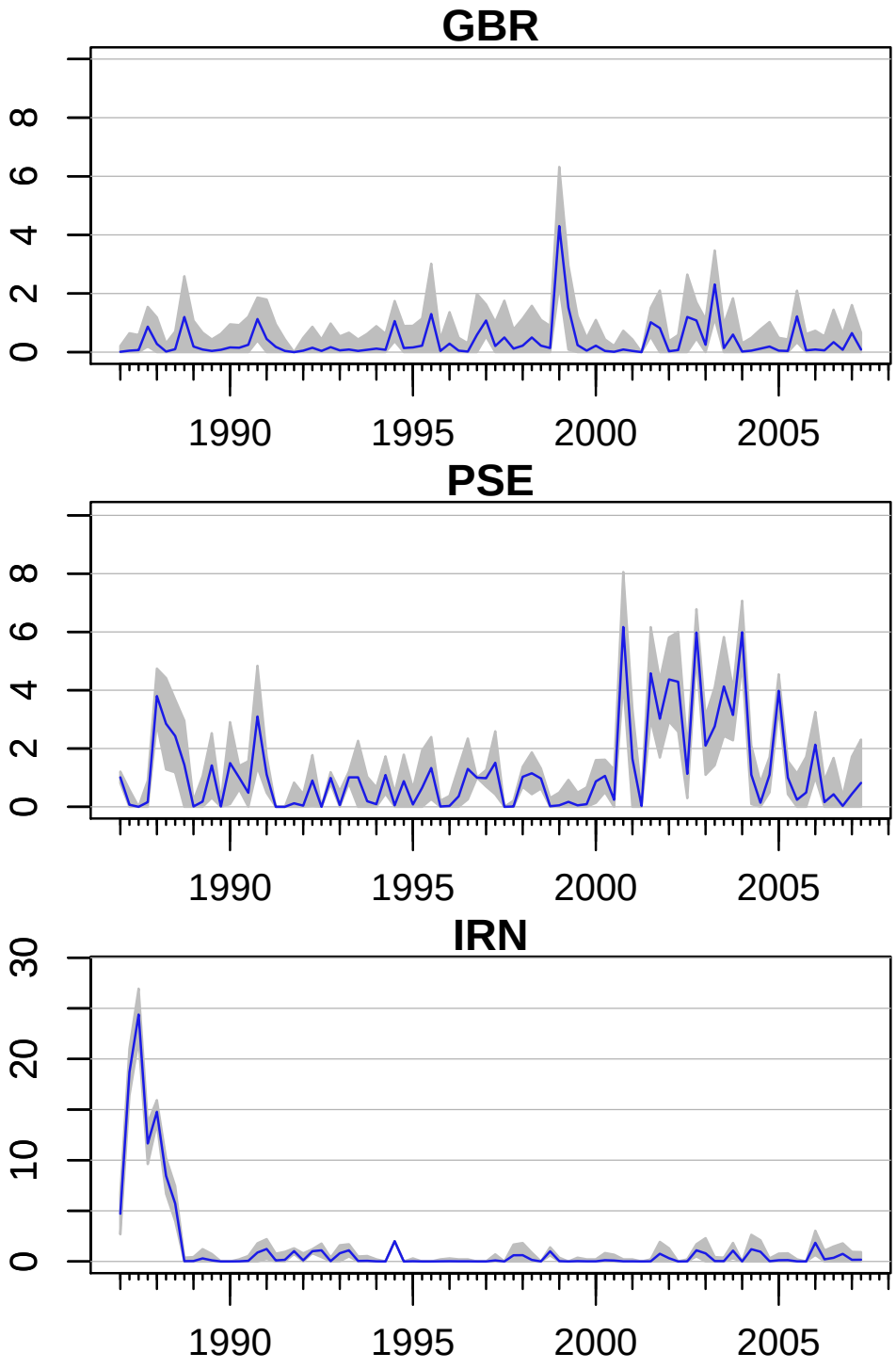
POS, NER, and constituent and dependency parses are produced with Stanford CoreNLP 3.5.2 with default settings except for one change, to use its shift-reduce constituent parser (for convenience of processing speed). We treat tags and parses as fixed and leave their uncertainty propagation for future work.

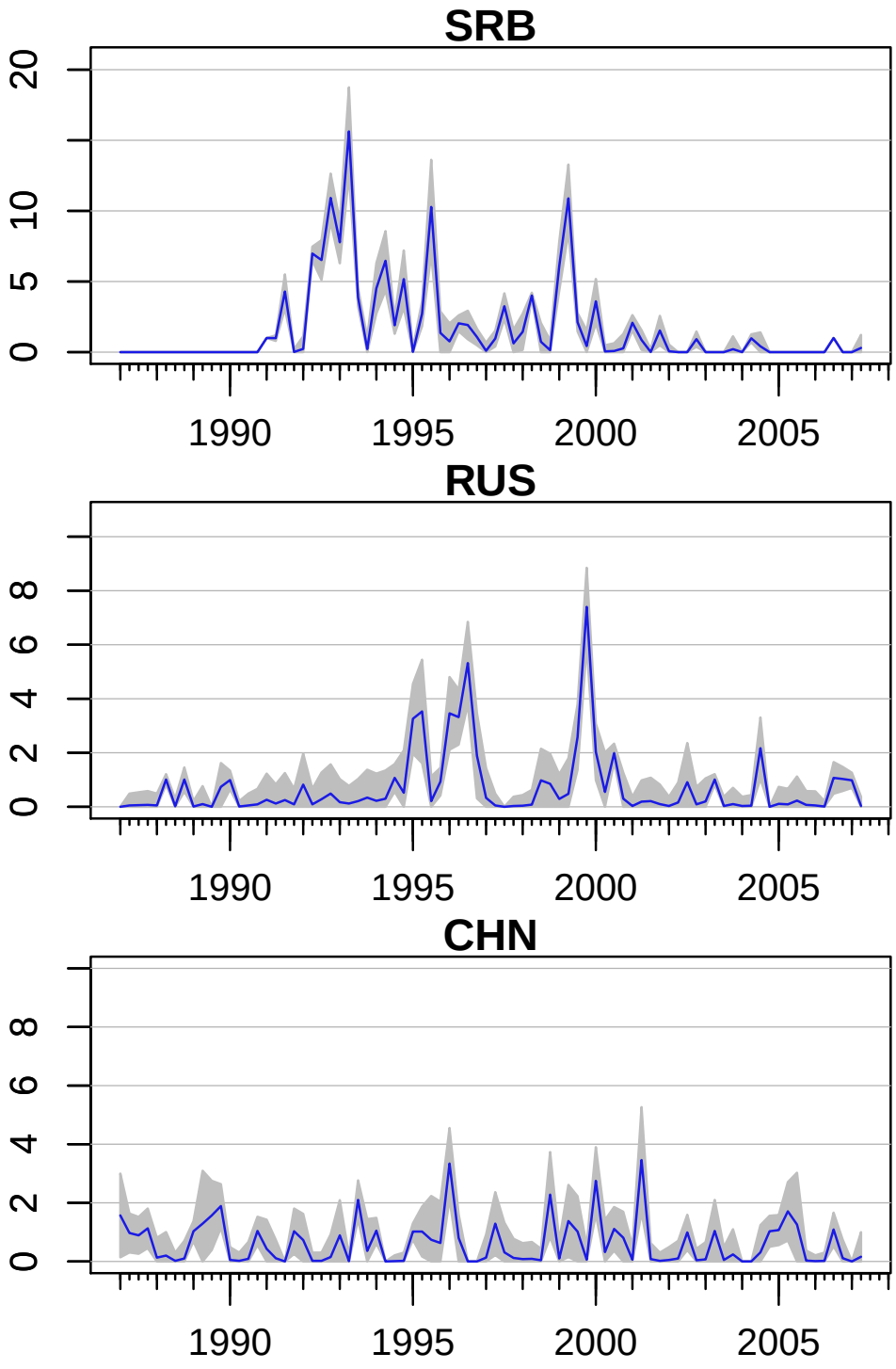
When formulating the extraction rules, we examined frequencies of all syntactic dependencies within country-affiliated entities, in order to help find reasonably high-coverage syntactic relations for the "attack" rule.

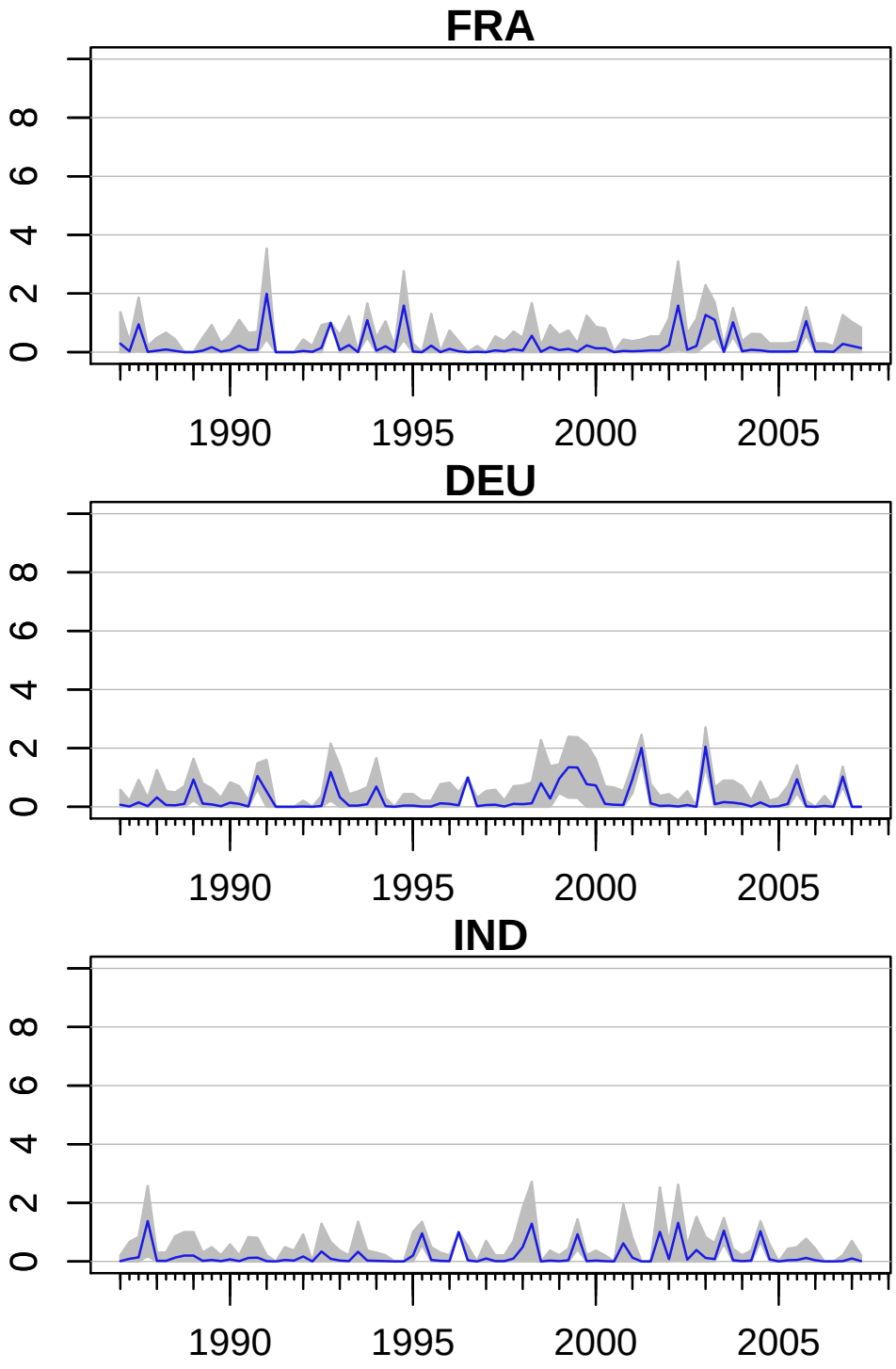
4 Event time series graphs

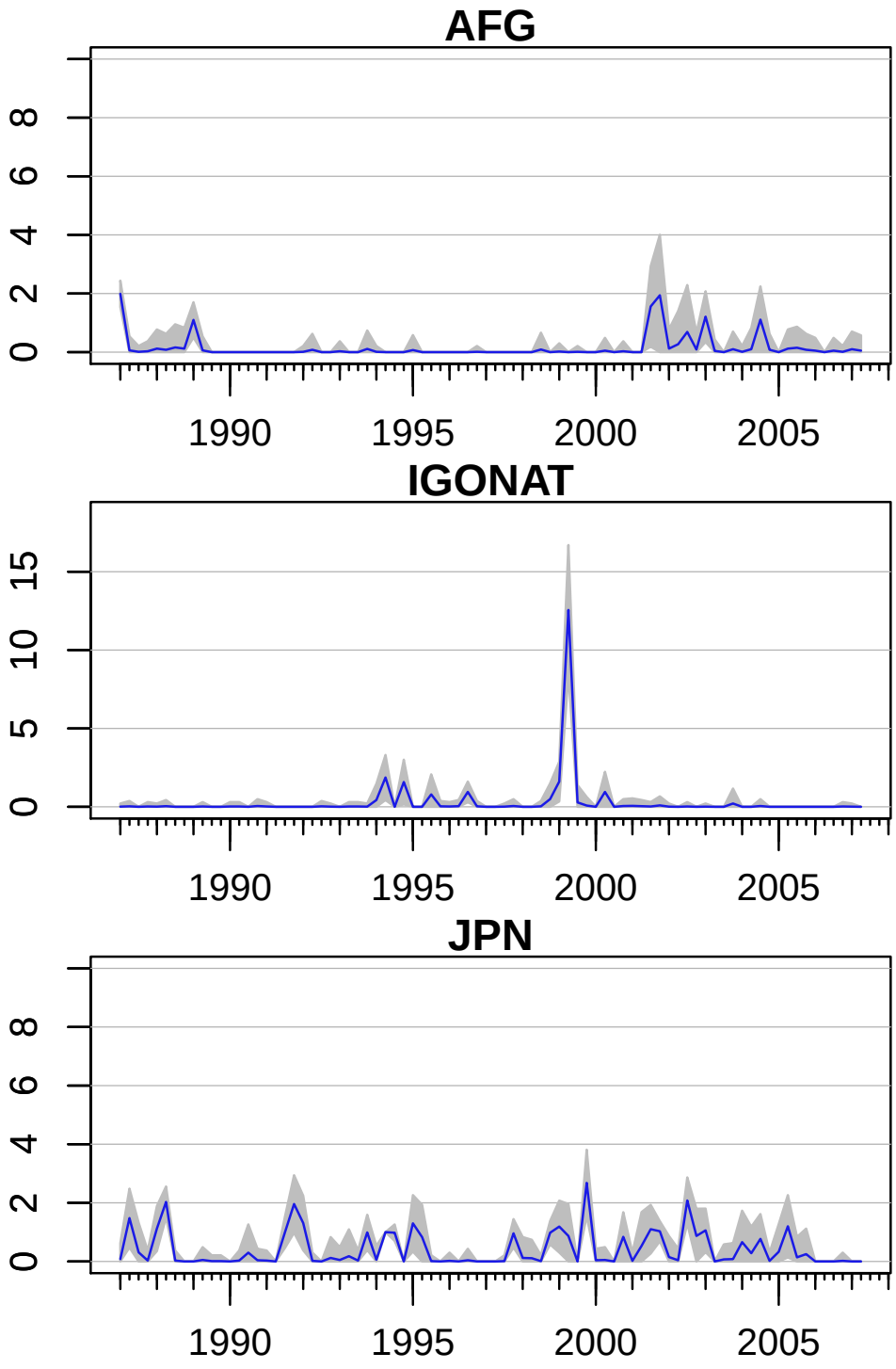
The following pages contain posterior time series graphs for 20 countries, as described in the section on coreference-based event aggregation, in order of decreasing total event frequency. As in the main paper, the blue line indicates the posterior mean, and the gray region indicates 95% posterior credibility intervals, with count aggregation at the monthly level. The titles are ISO3 country codes.

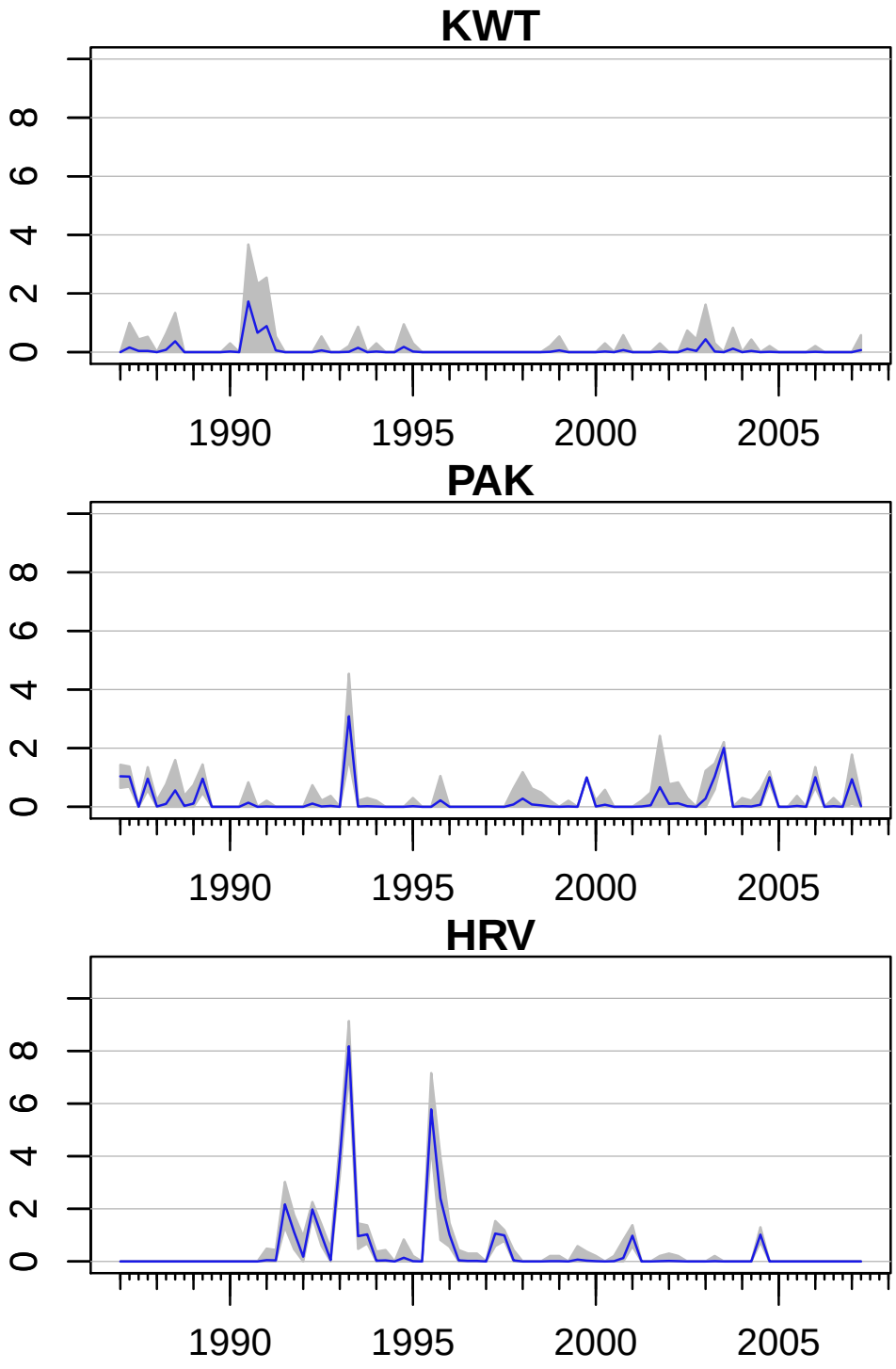

USA
0
10
20
30
1990
1995
2000
2005
IRQ
0
10
20
30
40
1990
1995
2000
2005
ISR
0
5
10
20
1990
1995
2000
2005

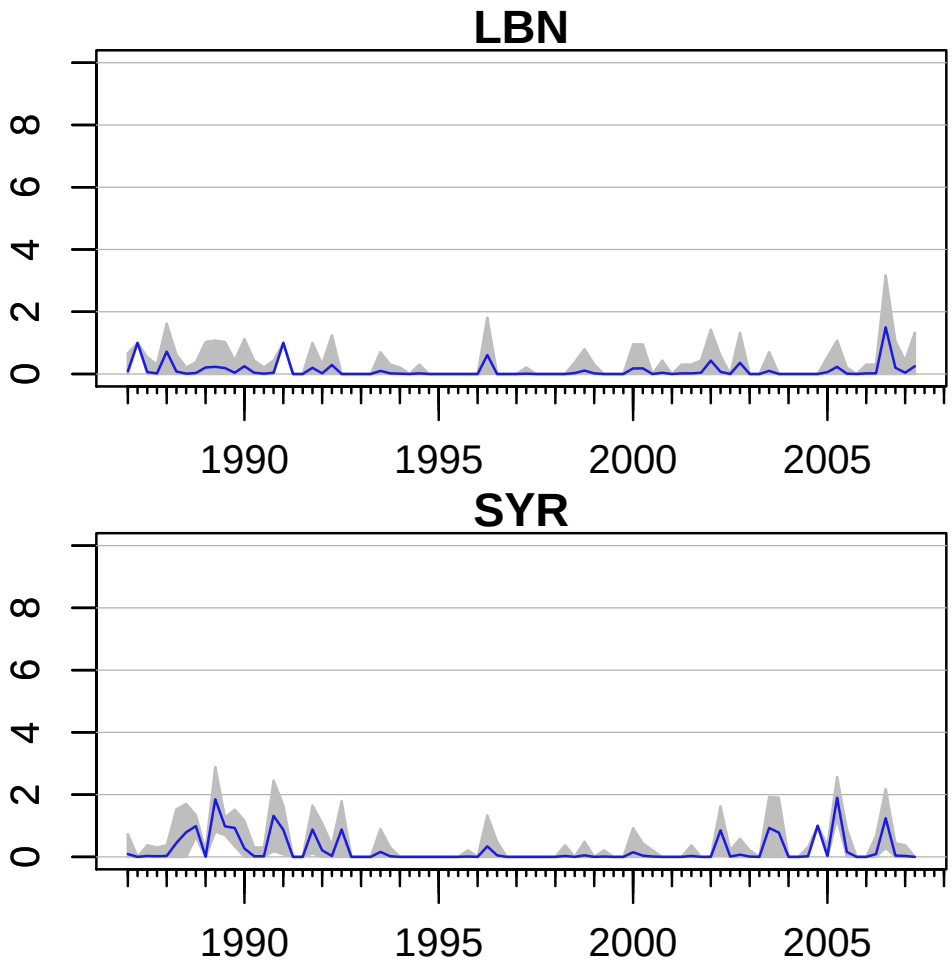

GBR
0
2
4
6
8
1990
1995
2000
2005
PSE
0
2
4
6
8
1990
1995
2000
2005
IRN
0
10
20
30
1990
1995
2000
2005


SRB
0
5
10
20
1990
1995
2000
2005
RUS
0
2
4
6
8
1990
1995
2000
2005
CHN
0
2
4
6
8
1990
1995
2000
2005


FRA
0
2
4
6
8
1990
1995
2000
2005
DEU
0
2
4
6
8
1990
1995
2000
2005
IND
0
2
4
6
8
1990
1995
2000
2005


AFG
0
2
4
6
8
1990
1995
2000
2005
IGONAT
0
5
10
15
1990
1995
2000
2005
JPN
0
2
4
6
8
1990
1995
2000
2005


KWT
0 2 4 6 8
1990
1995
2000
2005
PAK
0 2 4 6 8
1990
1995
2000
2005
HRV
0 2 4 6 8
1990
1995
2000
2005


LBN
0
2
4
6
8
1990
1995
2000
2005
SYR
0
2
4
6
8
1990
1995
2000
2005