IDENTIFYING TIME ZONES IN A LARGE DATASET OF MUSIC LISTENING LOGS

Gabriel Vigliensoni
gabriel@music.mcgill.ca

Ichiro Fujinaga
ich@music.mcgill.ca

AIMS

- Finding typologies of listeners by analyzing their music listening behaviour
- Identifying the time zone where a large amount of people have been listening to music, given their music listening histories

DATASET COLLECTION

- 594K music listening histories collected from last.fm
- last.fm API's `user.getRecentTracks()` → `<username, timestamp, artist-MBID, album-MBID, track-MBID>`
- Listeners filtered by
  - Submitting music logs for at least two years
  - Average activity of ten music logs per day
- Data cleaning by filtering out:
  - duplicated music logs
  - music logs which were less than 30 seconds apart in time

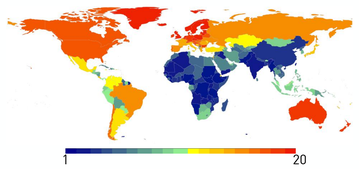

Figure 1: Last.fm penetration per country vigintiles

	Listeners	Logs	Artists	Albums	Tracks
	594K	27MM	555K	900K	7M
	Min	**1st Q**	**Median**	**3rd Q**	**Max**
Age	0	21	24	27	113
No. logs	7K	24K	37K	60K	998K
Life (days)	731	1192	1653	2188	3929

Table 1: Dataset demographics

TIME ZONE IDENTIFICATION

Time zone 0 cross correlation approach

- Approach based on the idea that listeners, in general, share a similar listening profile
- Calculated the lag value *k* which returned maximum correlation between `x[t+k]` and `y[t]` given a cross correlation function `ccf(x,y)`
- As fixed control time series `y[t]` we chose weekly aggregated listening profiles for 42K listeners in the UK (TZ 0)
- Listening behaviour:
  - no differences between weekdays and weekends
  - differences between day and night

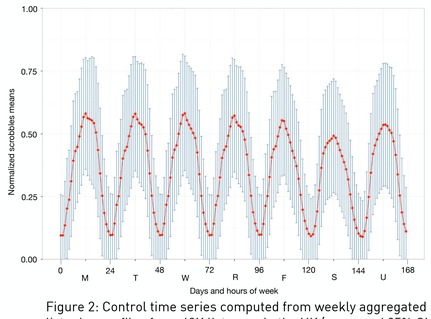

Figure 2: Control time series computed from weekly aggregated listening profiles from 42K listeners in the UK (mean and 95% CI error bars)

Local minima approach

- Approach based on the idea that listeners, in general, sleep during night time and submit fewer music logs
- We looked for the multi-local minima within a week, but for weekdays only
- Time zone computed by rounding the average of the normalized local minima values in the [-12, 11] range
- A variant based on averaging the forward and the backward computation of the local minima was implemented to calculate the middle point of "flat zones"

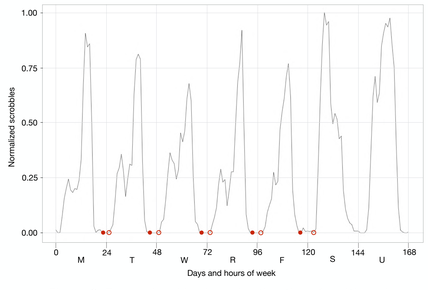

Figure 3: Forward and backward estimation of multi local minima of a weekly aggregated listening profile

Seasonal decomposition

- We employed time series decomposition to check if isolating cyclic seasonal data from any trend or noise in the weekly aggregated listening profiles would improve the performance of the time zone identification approaches

EXPERIMENT

Control dataset

- Created from 384 random listening histories from listeners in the dataset
- Aggregated their data into weekly listening profiles
- Manually labelled each one of these profiles in a time zone
- To evaluate the performance of each method by calculating the percentage of time differences between their computed time zone and the ones in the control dataset
- 1,000 populations replicated from the control dataset using bootstrapping method (resampling)

Compared approaches

	Time zone 0 cross correlation	Local minima-based approaches	
		Forward only	Forward-Backward average
Raw method	TZ0_XCORR	FF_LM	FB_LM
Seasonally decomposed	SEAS_TZ0_XCORR	SEAS_FF_LM	SEAS_FB_LM

Table 2: Six approaches for time zone identification and variants

Results

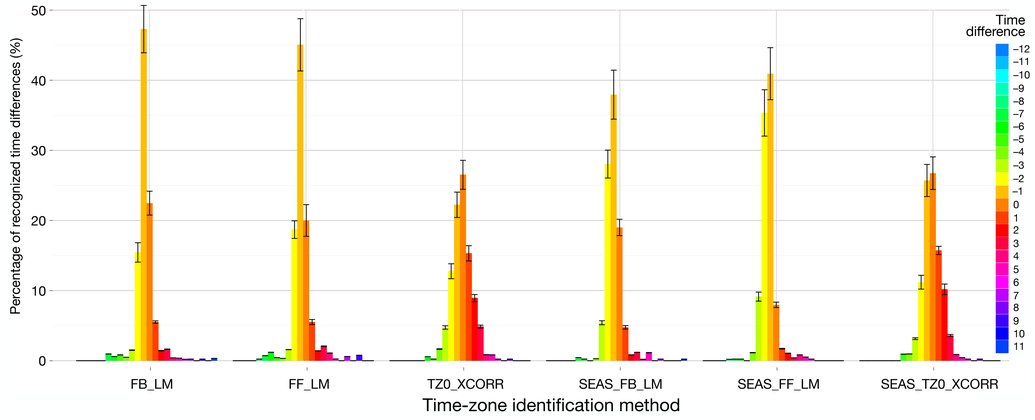

Figure 4: Performance of six approaches for time zone identification of listening profiles. The plots are shown with error bars indicating 95% confidence interval for each time difference between manually labelled and computed time zones for 1,000 populations taken with replacement from a sample of 384 random listening histories

CONCLUSIONS

- Time zone 0 cross-correlation approach yielded the best absolute performance
- However, the local minima approaches yielded a better performance if considering a tolerance window of ±1 hour
- The FB_LM variant for computing the middle point of "flat zones" did not improve the results substantially
- The seasonally decomposed versions had a poorer performance, which might imply that we are loosing information relevant for time zone identification
- Although individual schedule variations could exceed differences in time zones, the aforementioned approaches can still be used to shift listening profiles in time, allowing a more straightforward comparison

Acknowledgements

This research has been supported by BecasChile Bicentenario, CONICYT, and the Social Sciences and Humanities Research Council of Canada. Important part of this work was made using ComputeCanada's High Performance Computing resources. The authors also would like to thank Eugenio Tacchini for his idea about calling the Last.fm API using IDs instead of usernames