Estimation Lift through linear modeling

Tatari Interview Presentation

Ivan E. Perez

Hunter College - CUNY

March 27, 2021

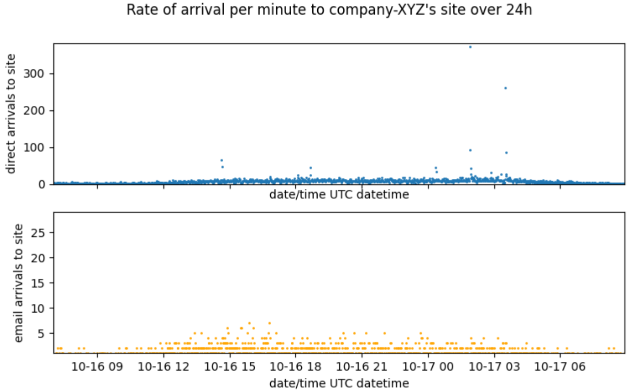
Presentation Outline

1. Background and Description of the Data
2. Point Processes
3. Identifying periods of influence from spot activity
4. Experiments and Results
5. performance metrics and actionable conclusions
6. future ideas and new metrics Do this if I have time
7. Suggested Improvements for Tatari Dashboard

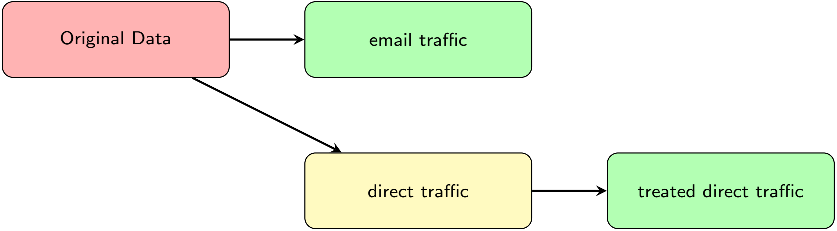
Description and treatment of data: Basic Description of Files

Web Traffic data

1. filename: `assignment-analyst-1-web-traffic-data.csv`
2. columns:
   - **time** as UTC datetime, that pandas converts to `pd.timeseries.datetime`
   - **source** as string, either `'direct'` or `'email'`
   - **value** as `int` and `float64`
3. time range: 10/16/2017 7:05pm to 11/13/2017 7:05 pm
4. web traffic range: **find this please**



Cleaning the Data I: Web Traffic Data



Original Data

1. time: pandas datetime
2. value: float64
3. traffic_source: str, email, direct

Split Data

1. email data
   a. time: pandas datetime
   b. value: int
2. direct traffic
   a. time: pandas datetime
   b. value: float

Imputed Direct Traffic **(didn't do)**

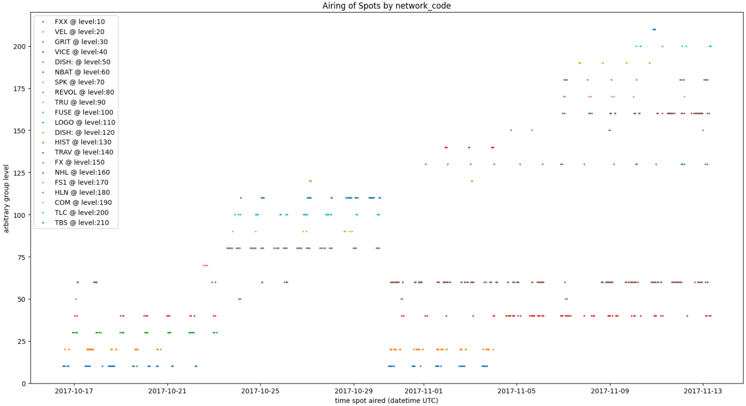
1. Negative reals replaced by ints
2. smoothing/grouping into 5-minute sections.

Cleaning the Data II: Spot Data Description

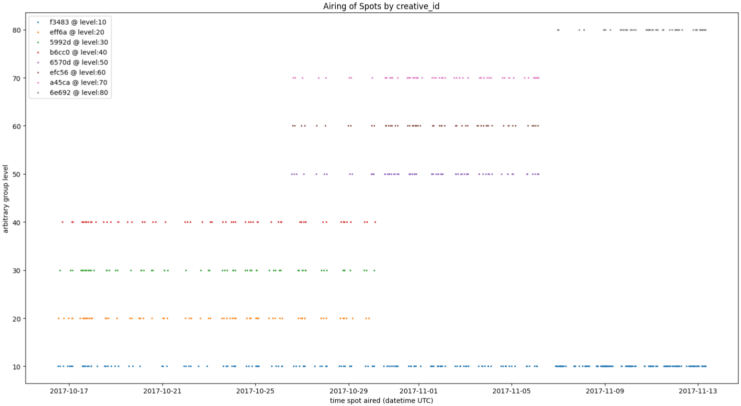
Column	Description	format
id	Spot identifier	integer
time	Time at which spot aired in local time with UTC offset	pd.datetime tz info
creative_id	Creative asset identifier	string
spend	Effective cost of spot USD	float64
program	Program during which spot aired	string
duration	Duration of spot in seconds	int, nan
network_code	Network on which the spot aired	string, nan
is_dual_feed	Specifies that the spot is aired at the same local time	True for all
rotation	Description of the target rotation where the spot aired	string, nan
rotation_days	Days of the week on which the roration applies to begining with monday=1	string, nan
rotation_start	Time when rotation starts	datetime
rotation_end	Time when rotation ends	datetime
feed	Which feed (East or West Coast) is the spot airing in	1, 2

1. filename: `assignment-analyst-1-spot-data.csv`
2. time range: 10/16/2017 7:05pm to 11/13/2017 7:05 pm

Exploring Spot Data: By Channel



Exploring Spot Data: By creative id



More categorical and numerical groupings available in figures folder.

Modeling Website traffic as a Poisson Point process:

Definition: time-Homogeneous Poisson Point Process

Consider the homogeneous Poisson counting process, $\{N(t)\}_{t\geq 0}$ with rate λt. It is said to be a *Poisson process* with rate $\lambda t > 0$ if it initializes at 0, has independent stationary increments and the following conditions hold:

(i) $\mathbb{P}(N(t+s) - N(t) = 1) = \lambda s + o(s)$

(ii) $\mathbb{P}(N(t+s) - N(t) \geq 2) = o(s)$

where $\{N(t)\}_{t\geq 0}$ is a Poisson process.

Key Assumptions:

1. Arrivals to the site in non-overlapping periods are independent from one another.
2. The rate of arrival is constant and uniform across any period.

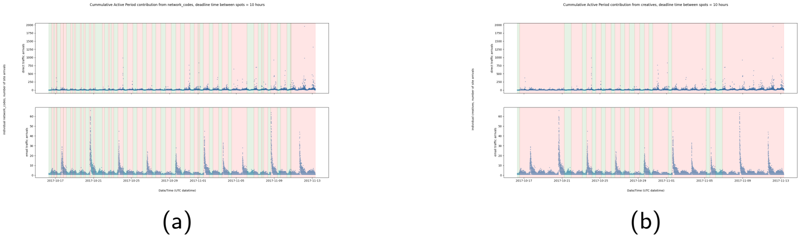
Goal 1: Identify periods where spots are played on each network to identify which networks lead to *increased* rates of website visits.
Goal 2: Identify periods where different versions of the same spots are played to identify which "creative id" leads to *increased* rates of website visits.

Goal 3: *Estimate lift-rate as the net increase in rate attributable to a spot, grouped by network or creative id.*

Identifying periods of influence from spot activity

An Informal[1] definition for the different periods from the set of all observed traffic visits, separated by traffic source we first define

1. Quiet Periods a the set of observed site visits when **No** spots not been played recently.
2. Active Periods for an identifier (e.g., network code, or creative id) as traffic when the spots associated with the identifier have been player recently.



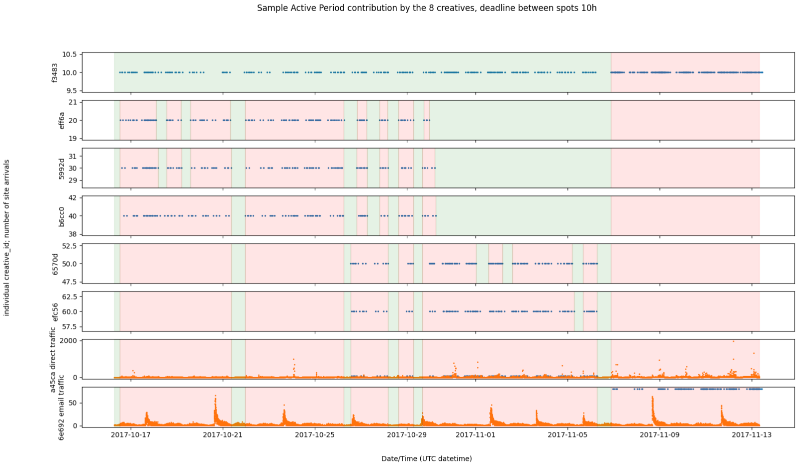
(a)

(b)

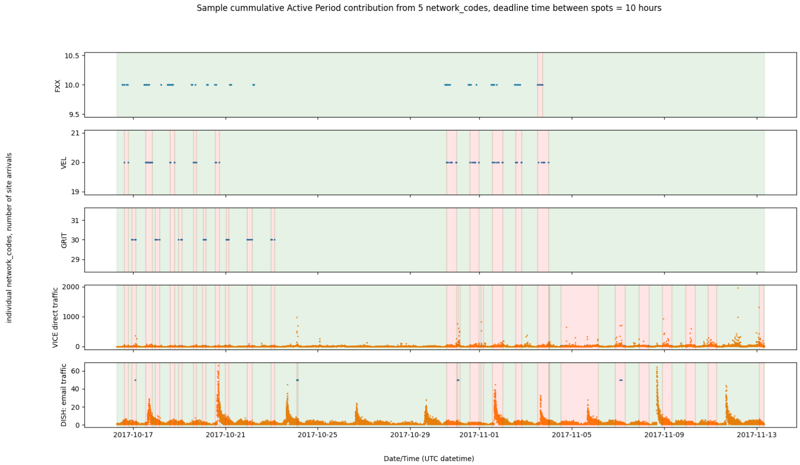
Figure: (a) Cumulative Active periods (Red) from spots grouped by (a)network code, (b) creative id. (Green) represents the quiet periods

[1] An incomplete definition can be found in the latex file

Showing how the each creative builds up the total active period



Showing how the each network code builds up the total active period



t-test to compare Active and Quiet Periods and its limitations

Rationale:

1. By identifying periods where no spots are being played, we can identify a base rate of arrival to the site through email, $\hat{\lambda}_0^{\text{eml}}$, and direct traffic $\hat{\lambda}_0^{\text{dir}}$.
2. By comparing to the individual active periods for each spot and their network code or creative id we can see if the presence of an spot on a network or individual creative id contributes significantly leads to an elevated rate of site arrivals.

Experimental Setup:

1. Let the identifier, n, have a N Active periods. The i^{th} Active period, describes the estimated rate of arrival, $\hat{\lambda}^n_{1,i}, i = 1, 2, \ldots, N$ in that period.
2. To compare the assumed independent populations of active periods, $\left\{\hat{\lambda}^n_{1,i}, \forall i \in N\right\}$, and M quiet periods, $\left\{\hat{\lambda}_{0,i}, \forall i \in M\right\}$, we employ a t-test.
3. We define population means for the quiet and Active periods as μ_0 and μ_1^n respectively.
4. Our test becomes: $\begin{aligned} H_0 : &\quad \mu_1^{\mathrm{n}} - \mu_0 = 0, \\ H_1 : &\quad \mu_1^{\mathrm{n}} - \mu_0 > 0 \end{aligned}$

Assumptions and Limitations:

- The periods are independent of one another.
- The statistic is less valid for low sample sizes, (which occurrs for spots that generate few periods)
- weaker but still present assumption of normality for the distribution of $\hat{\lambda}$'s.

t-test results: email data by Network

channel	μ_1 visits/min	t-calc	t-crit @ $\alpha = 0.05$	Reject H_0?	Lift, $\mu_1 - \mu_0$	Total Spend (USD)
FXX	2.3921	2.1721	1.6790	YES	1.3524	6477.00
VEL	4.4217	3.6548	1.6820	YES	3.3820	11152.00
GRIT	1.6615	1.0509	1.6840	NO	0.6218	4760.00
VICE	1.5247	1.1486	1.6780	NO	0.4850	15317.00
DISH:ESPN	1.5616	0.5921	1.6880	NO	0.5219	17212.50
NBAT	3.2167	3.6821	1.6770	YES	2.1770	17902.70
SPK	0.9825	-0.0378	1.6910	NO	0.0572	2890.00
REVOLT	2.2442	1.8805	1.6840	YES	1.2045	2210.00
TRU	1.4554	0.3858	1.6900	NO	0.4157	3740.00
FUSE	1.5391	0.8507	1.6840	NO	0.4994	3403.40
LOGO	1.7982	1.0516	1.6860	NO	0.7585	5397.50
DISH:NFLN	1.2984	0.2421	1.6900	NO	0.2587	3060.00
HIST	1.7743	0.9667	1.6870	NO	0.7346	23247.50
TRAV	2.4299	1.5684	1.6880	NO	1.3902	5100.00
FX	3.8182	1.8388	1.6910	YES	2.7785	10540.00
NHL	1.5646	0.7613	1.6860	NO	0.5249	7352.50
FS1	1.6658	0.7050	1.6880	NO	0.6261	3179.00
HLN	1.8803	0.9537	1.6880	NO	0.8406	8160.00
COM	14.0825	3.3233	1.6870	YES	13.0428	2550.00
TLC	1.1306	0.1031	1.6880	NO	0.0909	6800.00
TBS	1.0467	0.0046	1.6910	NO	0.0070	5100.00

t-test results: direct data by Network

channel	μ_1	t-calc	t-crit @	Reject	Lift,	Total
	visits/min		$\alpha = 0.05$	H_0?	$\mu_1 - \mu_0$	Spend (USD)
FXX	8.9825	1.1305	1.6790	NO	1.5840	6477.00
VEL	12.4398	3.4039	1.6820	YES	5.0413	11152.00
GRIT	9.9908	1.5448	1.6840	NO	2.5923	4760.00
VICE	22.1660	7.1580	1.6780	YES	14.7675	15317.00
DISH:ESPN	48.1157	11.5616	1.6880	YES	40.7172	17212.50
NBAT	24.6751	8.9567	1.6770	YES	17.2766	17902.70
SPK	8.5671	0.2634	1.6910	NO	1.1686	2890.00
REVOLT	14.0415	3.8185	1.6840	YES	6.6430	2210.00
TRU	11.5695	1.3287	1.6900	NO	4.1710	3740.00
FUSE	15.6779	4.3325	1.6840	YES	8.2794	3403.40
LOGO	14.4085	3.3187	1.6860	YES	7.0100	5397.50
DISH:NFLN	24.7257	5.2490	1.6900	YES	17.3272	3060.00
HIST	46.4140	10.9205	1.6870	YES	39.0155	23247.50
TRAV	21.4131	3.8085	1.6880	YES	14.0146	5100.00
FX	59.8364	11.8217	1.6910	YES	52.4379	10540.00
NHL	23.8175	8.3011	1.6860	YES	16.4190	7352.50
FS1	24.8399	6.5103	1.6880	YES	17.4414	3179.00
HLN	43.4580	13.9879	1.6880	YES	36.0595	8160.00
COM	26.1737	7.1864	1.6870	YES	18.7752	2550.00
TLC	27.9236	7.8521	1.6880	YES	20.5251	6800.00
TBS	24.1560	3.7778	1.6910	YES	16.7575	5100.00

t-test results by creative id

Email Results

creative_id	μ_1 visits/min	t-calc	t-crit @ $\alpha = 0.05$	Reject H_0?	Lift, $\mu_1 - \mu_0$	Total Spend (USD)
f3483	2.3608	1.4911	1.6960	NO	1.7517	70555.1000
eff6a	2.0483	1.2591	1.7140	NO	1.4392	11391.7000
5992d	2.0586	1.3143	1.7140	NO	1.4496	11099.3000
b6cc0	2.2467	1.4401	1.7170	NO	1.6376	10897.8500
6570d	2.4830	1.9204	1.7210	YES	1.8739	12630.1500
efc56	2.3951	2.1186	1.7210	YES	1.7860	8749.9000
a45ca	2.3823	2.2615	1.7290	YES	1.7732	9520.8500
6e692	1.7295	0.5922	1.7400	NO	1.1204	30706.2500

Direct Results

creative_id	μ_1 visits/min	t-calc	t-crit @ $\alpha = 0.05$	Reject H_0?	Lift, $\mu_1 - \mu_0$	Total Spend (USD)
f3483f	13.8871	0.3346	1.6960	NO	8.6028	70555.10
eff6a	11.5753	0.3095	1.7140	NO	6.2911	11391.70
5992d	11.3258	0.3041	1.7140	NO	6.0415	11099.30
b6cc0	12.1324	0.3522	1.7170	NO	6.8482	10897.85
6570d	15.1600	0.4465	1.7210	NO	9.8757	12630.15
efc56	14.9599	0.4435	1.7210	NO	9.6756	8749.90
a45ca	14.2705	0.5527	1.7290	NO	8.9863	9520.85
'6e692	25.6538	0.9468	1.7400	NO	20.3696	30706.25

Actionable Conclusions

Email traffic conclusions:
We determined that the networks where spots contributed to increased traffic were, FXX, VEL, NBAT, REVOLT, FX, COM[2]. We determined that the creatives that were better received began with 6570d, efc56, and a45ca.

Direct traffic conclusions:
Many networks showed increased rates of site traffic, but many were erratic, unverifiable and short lived. Without further understanding of direct traffic data, I cannot make more effective conclusions beyond, the ineffective networks were GRIT, SPK, TRU.

Measuring Cost Effectiveness of our Spots:
These metrics
would have been good to calculate, it can be done by counting net total site visits instead of a rate.

Metric Name	Abbrev.	Description	Formula
Cost per Visit	CPV	Amount Spent per Visitor	Total cost of spots in all active periods/ Total No. of visits in the periods
Spot Effectiveness	SE	How often a spot must be shown to get increased traffic	Total No. of visits / No. times the spot was shown
Network Effectiveness	CE	How many channels a spot must be shown on to get increased traffic	No. of networks where the spot is shown/ Total No. of visits across all networks

[2] low counts for active periods.

Future Idea: Modeling arrivals Hawke's processes

Rationale:
On slides 9,10,11 we see that when there is an engaging spot campaign, there is an initial jump with a long drift back to baseline. This implies that spot campaigns have a lasting effect, violating the independence assumption of simple Poisson point processes. To capture that we introduce Hawke's Processes.

Definition: Hawkes Process[a]

[a] https://en.wikipedia.org/wiki/Point_process

A Hawkes process $\{Q(t)\}_{t\geq 0}$, also known as a self-exciting counting process, is a simple point process whose conditional intensity can be expressed as:

$$\lambda(t) = \mu(t) + \int_{-\inf}^{t} \nu(t-s)dN(s),$$

where $\nu : \mathbb{R}^+ \to \mathbb{R}^+$ is a kernel function (i.e., $\nu(x) = \alpha e^{-\beta x}$) which expresses the positive influence of past events on the current rate process, $\lambda(t), \mu(t)$.

Idea: Hawke's parameter estimation $\left\{\hat{\alpha}, \hat{\beta}, \hat{\mu}, \right\}$ through stochastic optimization and simulation[3].

[3] Da Fonseca, J., & Zaatour, R. (2014). Hawkes process: Fast calibration, application to trade clustering, and diffusive limit. Journal of Futures Markets, 34(6), 548-579

Suggested Improvements for Tatari Dashboard

1. User Interface:
   - In the linear panel the zoom could be adjusted by stretching the window instead of by a slider.
   - Users could have the option of adjusting granularity through a drop-down menu.
   - Toggle switch for axes sharing between top and bottom graphs in linear
   - Allowing users to change the date window.
2. Features:
   - in calendar view for CPV, include an estimate for what the media slot has sold for in the past, to illustrate savings, or premium paid.
3. Issues:
   - Creative heat map disabled and makes it hard to click back.
   - I'm not sure why the bar graphs for spots, limits colors to two channels each, may just be an glitch on my end.