Using Browser Data to Understand Desires to Spend Time Online

Supplementary Material

Supplementary Methods

Additional Measures

Participants indicated the locations for their internet use by selecting all applicable options from those presented ("At Home", "At Work", "At School", "In Public", or "Other Distinct Setting"). If "Other Distinct Setting" was selected they were prompted to name the setting. Whether frequent internet use was a school or work requirement was indicated with a binary response (Yes/No). The fourth item asked them to rate their skill/knowledge for using the internet on a 7-point scale ranging from "Less Skilled/Knowledgeable" (1) to "More Skilled/Knowledgeable" (7). Finally, for each of the internet usage locations selected, participants were asked to indicate, using a slide ranging from "Wasted" through "either equally" to "well spent", whether certain types of internet use were viewed as wasteful or time well spent.

Synthetic Data

To produce a synthetic dataset to enable us to share our proprietary and potentially sensitive data (see Quintana, 2020 for a primer) we used the *synthpop* R package (version 1.7; Nowok et al., 2016) in the R statistical environment. This dataset was produced with permission from the Mozilla Corporation.

A Synthetic dataset is designed to mimic the original dataset by maintaining its statistical properties and the relationships between variables, but without any synthetic record matching any of the values in the original dataset. Specifically, we used the default synthesis

method in *synthpop*, the classification and regression tree (CART) procedure (Reiter, 2005).[1] As Nowok et al. (2016) explain, the synthesis procedure with the CART model involves three steps:

1. Fit a classification or regression tree by binary recursive partitioning.
2. For each variable find the terminal node.
3. Randomly draw a donor from the members of the node and take the observed value of *y* from that draw as the synthetic value.

These steps were implemented with the *syn*() function with all parameters set to their defaults (see the code provided below). For the *minbucket* parameter, given recommendations for a minimum value of 20, we used 21 as the minimum number of observations in any terminal node of a CART model.

Following generation of the synthetic data, we used the *sdc*() function to identify and, if needed, remove any synthetic records that also appeared in the original dataset. No synthetic records matched those from the original dataset. This function also adds a label to the synthetic dataset to explicitly designate the data as 'fake'.

References:

Nowok B, Raab GM, Dibben C. 2016. Synthpop:bespoke creation of synthetic data in R. *Journal of Statistical Software* 74:1–26. DOI: https://doi.org/10.18637/jss.v074.i11.

Quintana, D. S. (2020). A synthetic dataset primer for the biobehavioural sciences to promote reproducibility and hypothesis generation. *Elife*, 9, e53275.

Reiter JP (2005). Using CART to Generate Partially Synthetic, Public Use Microdata. *Journal of Official Statistics*, 21(3), 441–462.

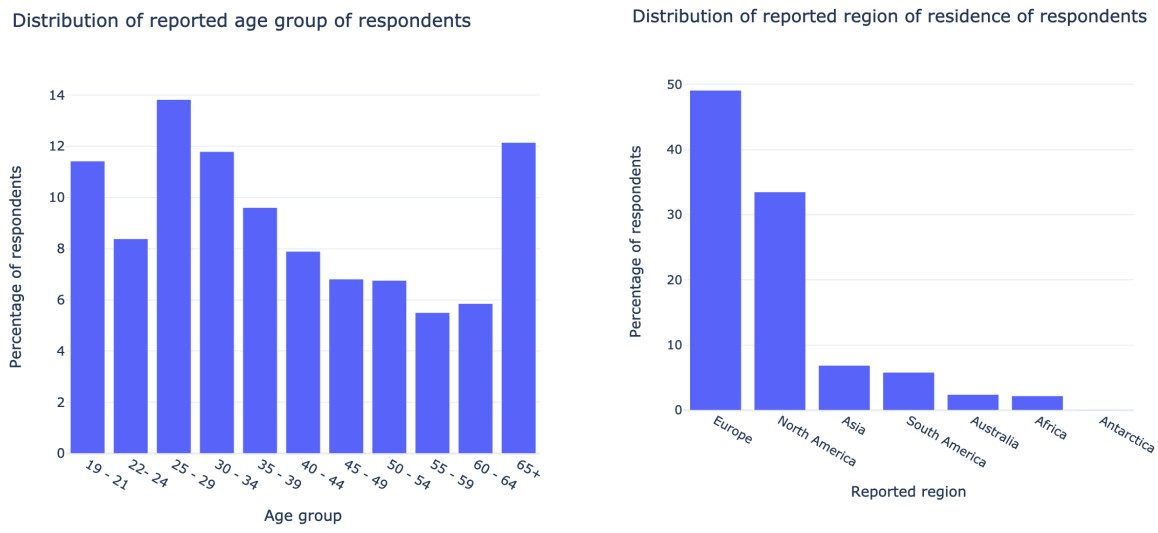
[1]See https://github.com/cran/synthpop/blob/master/R/syn.r for the implementation of this function.

Code for the production of a synthetic dataset

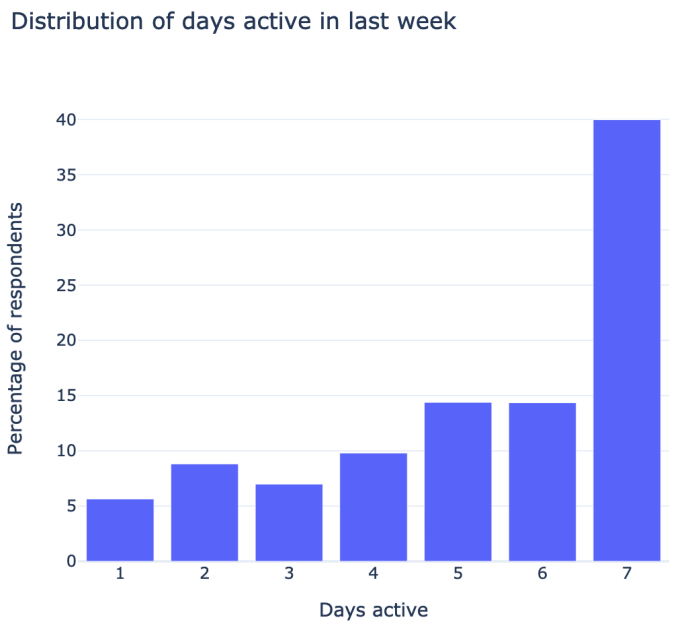
We used the following code to produce the synthetic dataset:

```
library(dplyr)
library(synthpop)
data <- read.csv("~/Downloads/gl_raw_data.csv")
data <- data %>%
  mutate(
    profile_age = as.Date(date) - as.Date(first_appearance_date),
    profile_at_least_week_old = profile_age >= 7
  ) %>%
  rename(
    days_active = is_active
  ) %>%
  select(
    age_group,
    region,
    days_active,
    hours,
    uri_count,
    domain_count,
    more_time,
    less_time,
    desktop_mobile,
    primary_browser,
    profile_at_least_week_old
  ) %>%
  mutate(
    age_group = as.factor(age_group),
    region = as.factor(region),
    more_time = as.factor(more_time),
    less_time = as.factor(less_time),
    primary_browser = as.factor(primary_browser),
    profile_at_least_week_old = as.factor(profile_at_least_week_old)
  )

syn_data <- syn(data, cart.minbucket=21)

sdc_data <- sdc(syn_data, data)

print(nrow(data))
print(nrow(syn_data$syn))
print(nrow(sdc_data$syn))
```

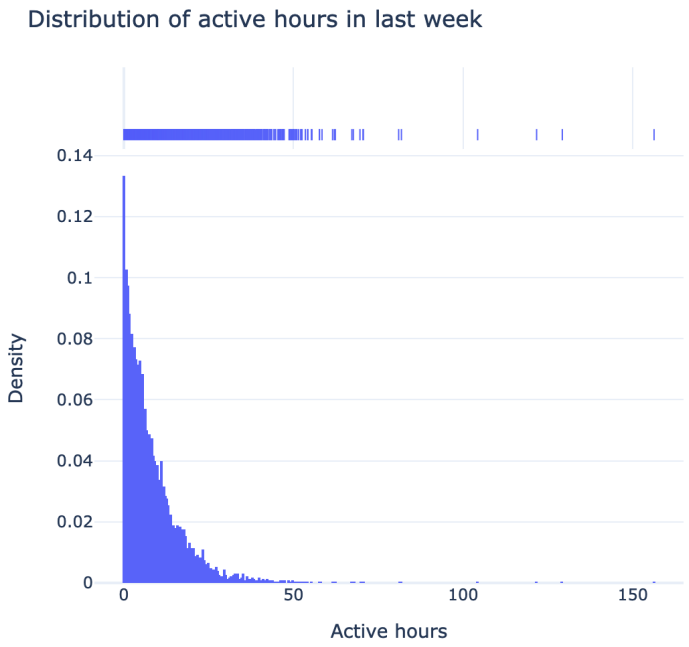
Supplementary Results



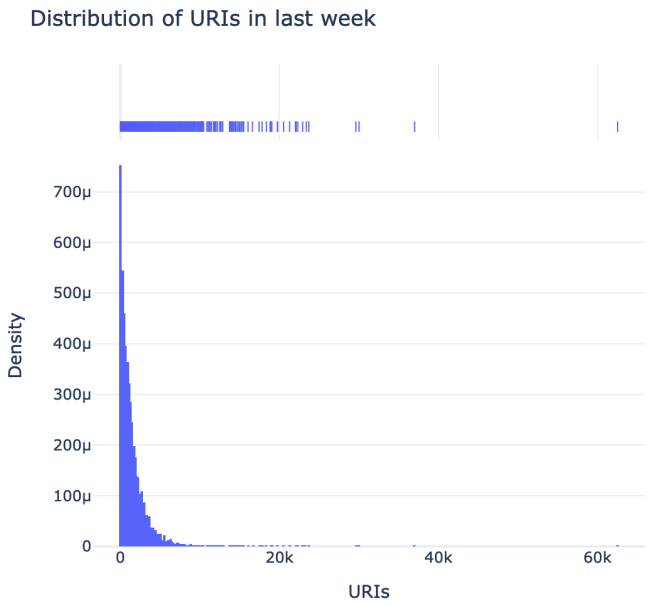
Supplementary Figure 1. Histograms for the demographic covariates: age group (excluding 2.77% of responses for which the variable is missing) and region of residence (excluding 2.88% of responses for which the variable is missing).



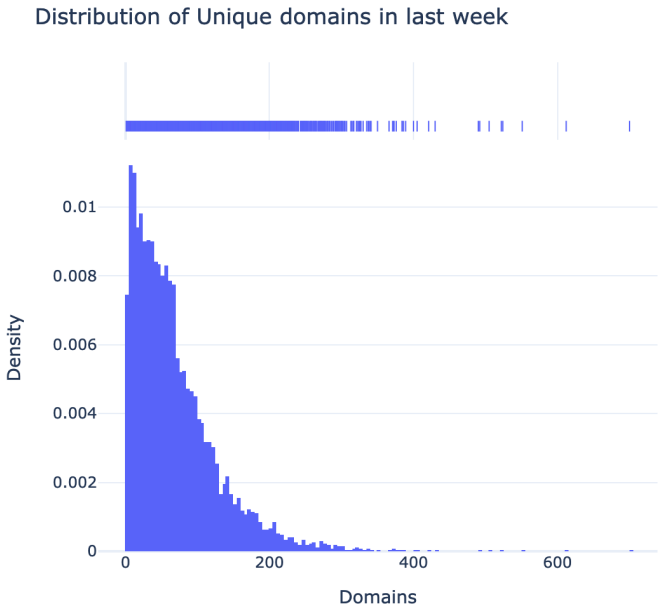
Supplementary Figure 2. Histogram for the days active variable.



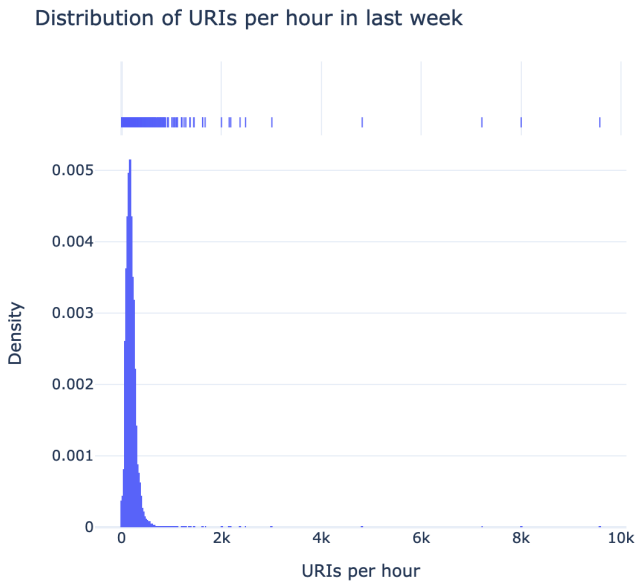
Supplementary Figure 3. Histogram with rug plot for the distribution of active hours in the last week.



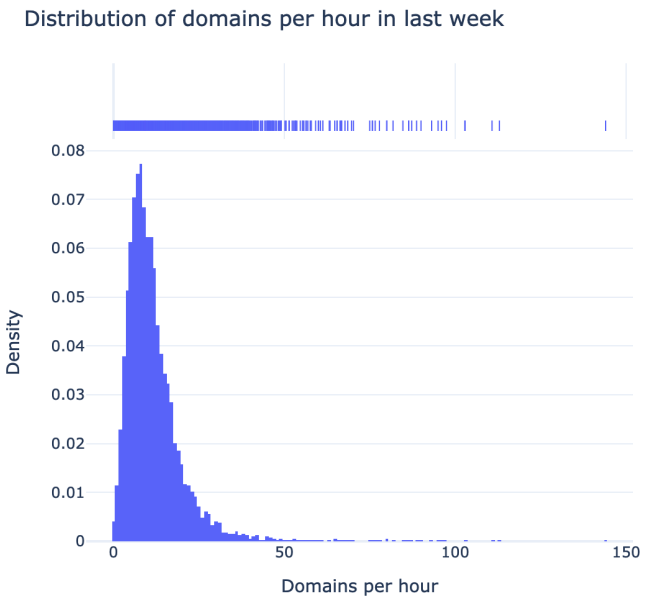
Supplementary Figure 4. Histogram with rug plot for the distribution of URIs in the last week.



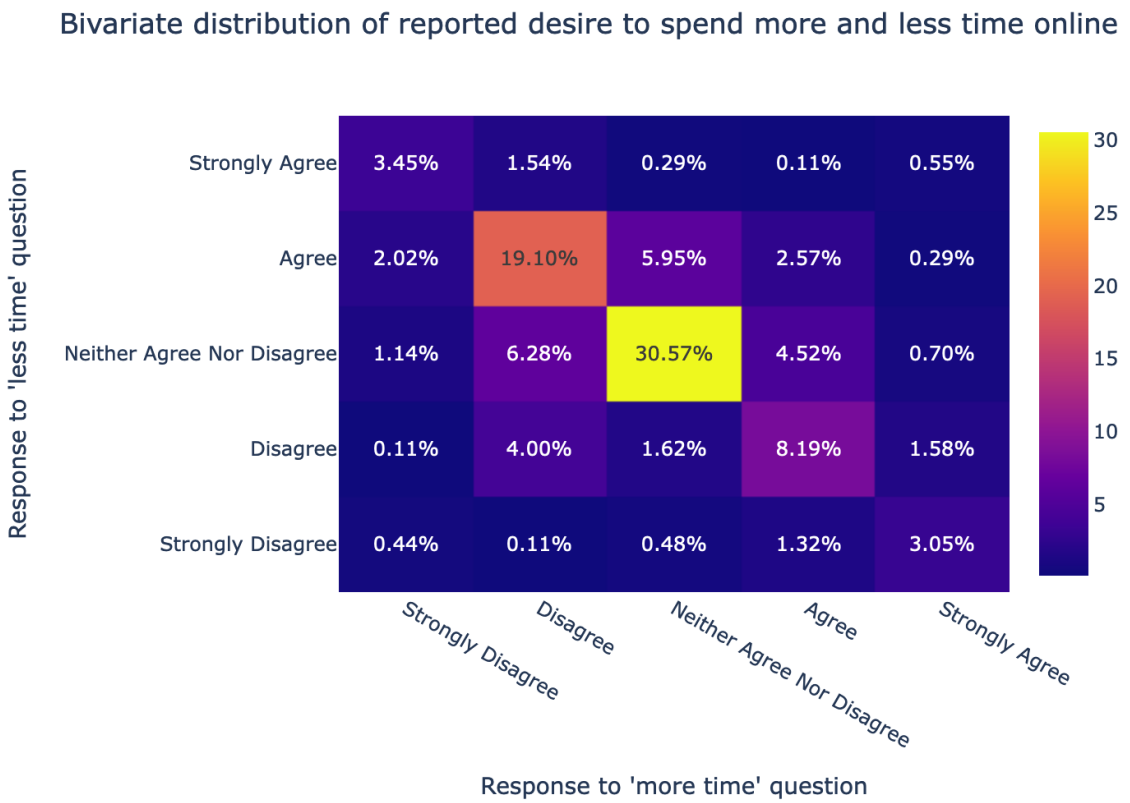
Supplementary Figure 5. Histogram with rug plot for the unique domains in the last week.



Supplementary Figure 6. Histogram with rug plot for the distribution of URIs per hour hours in the last week.



Supplementary Figure 7. Histogram with rug plot for the distribution of domains per active hour in the last week.



Supplementary Figure 8. Heatmap of the bivariate distribution of reported desire to spend more and less time online

Table S1. Frequencies for each age category.

Age category	Frequency (%)
25 - 29	13.44
65+	11.81
30 - 34	11.46
19 - 21	11.11
25 - 39	9.34
22 - 24	8.15
40 - 44	7.67
45 - 49	6.62
50 - 54	6.57
60 - 64	5.70
55 - 59	5.35
NA	2.76
Total	100

Table S2. Results of six separate ordinal logistic regressions predicting internet usage desire, while controlling for region and age group.

Variable	Standardised Estimate	p-val	95% CI	
			LL	UL
Active Days	0.096	0.00	0.055	0.137
Active Hours	0.046	0.026	0.005	0.086
URI Count	0.009	0.664	-0.032	0.050
Domain Count	-0.005	0.820	-0.045	0.036
URI per hour	-0.024	0.237	-0.063	0.016
Domain per hour	-0.049	0.017	-0.090	-0.009

Table S3. Descriptive statistics for the key study variables after accounting for outliers.

Variable	Outliers removed		
	n	M (SD)	Median
Active hours	7671	7.13 (5.92)	5.57
URI count	7535	1157.74 (995.15)	888.00
Domain count	7794	61.07 (45.66)	51.00
URI per hour	7840	179.09 (79.38)	171.52
Domains per hour	7710	10.37 (5.59)	9.43

Table S4. Results for age interactions

Variable	treatment	comparison	coef	std err	z	p-val	LC	UC	BF sig
Active days	25-29	19 - 21	-0.1138	0.087	-1.307	0.191	-0.285	0.057	
Active days	25-29	22 - 24	-0.1221	0.096	-1.267	0.205	-0.311	0.067	
Active days	25-29	30 - 34	-0.0811	0.088	-0.924	0.356	-0.253	0.091	
Active days	25-29	35 - 39	0.0425	0.091	0.465	0.642	-0.137	0.222	
Active days	25-29	40- 44	-0.1257	0.094	-1.335	0.182	-0.31	0.059	
Active days	25-29	45 - 49	-0.0741	0.096	-0.774	0.439	-0.262	0.114	
Active days	25-29	50 - 54	-0.0991	0.097	-1.026	0.305	-0.288	0.09	
Active days	25-29	55 - 59	-0.1783	0.102	-1.749	0.08	-0.378	0.022	
Active days	25-29	60 - 64	-0.1509	0.099	-1.517	0.129	-0.346	0.044	
Active days	25-29	65+	-0.1448	0.081	-1.777	0.076	-0.304	0.015	
Active hours	25-29	19 - 21	-0.1442	0.08	-1.794	0.073	-0.302	0.013	
Active hours	25-29	22 - 24	-0.0627	0.095	-0.662	0.508	-0.248	0.123	
Active hours	25-29	30 - 34	-0.0791	0.084	-0.94	0.347	-0.244	0.086	
Active hours	25-29	35 - 39	0.0318	0.089	0.359	0.72	-0.142	0.206	
Active hours	25-29	40- 44	-0.141	0.103	-1.373	0.17	-0.342	0.06	
Active hours	25-29	45 - 49	-0.0378	0.098	-0.386	0.7	-0.23	0.154	
Active hours	25-29	50 - 54	-0.0862	0.095	-0.904	0.366	-0.273	0.101	
Active hours	25-29	55 - 59	0.0846	0.105	0.802	0.423	-0.122	0.291	
Active hours	25-29	60 - 64	-0.1146	0.085	-1.352	0.176	-0.281	0.052	
Active hours	25-29	65+	-0.0306	0.077	-0.397	0.692	-0.182	0.121	
URI count	25-29	19 - 21	-0.1495	0.087	-1.723	0.085	-0.32	0.021	
URI count	25-29	22 - 24	-0.0983	0.085	-1.157	0.247	-0.265	0.068	
URI count	25-29	30 - 34	-0.1312	0.087	-1.509	0.131	-0.302	0.039	
URI count	25-29	35 - 39	-0.083	0.093	-0.895	0.371	-0.265	0.099	
URI count	25-29	40- 44	-0.0239	0.074	-0.325	0.745	-0.168	0.12	
URI count	25-29	45 - 49	-0.2594	0.105	-2.473	**0.013**	-0.465	-0.054	
URI count	25-29	50 - 54	-0.0484	0.1	-0.486	0.627	-0.243	0.147	
URI count	25-29	55 - 59	0.1536	0.126	1.215	0.224	-0.094	0.401	
URI count	25-29	60 - 64	-0.1285	0.133	-0.969	0.332	-0.388	0.131	
URI count	25-29	65+	-0.0012	0.078	-0.016	0.987	-0.154	0.151	
Domain count	25-29	19 - 21	-0.2963	0.087	-3.398	**0.001**	-0.467	-0.125	Yes
Domain count	25-29	22 - 24	-0.0987	0.096	-1.031	0.302	-0.286	0.089	
Domain count	25-29	30 - 34	-0.1927	0.081	-2.382	**0.017**	-0.351	-0.034	
Domain count	25-29	35 - 39	-0.1511	0.086	-1.759	**0.079**	-0.319	0.017	Yes
Domain count	25-29	40- 44	-0.2621	0.089	-2.939	**0.003**	-0.437	-0.087	Yes
Domain count	25-29	45 - 49	-0.2872	0.097	-2.971	**0.003**	-0.477	-0.098	Yes

Domain count	25-29	50 - 54	-0.0905	0.092	-0.985	0.325	-0.271	0.09	
Domain count	25-29	55 - 59	-0.0634	0.105	-0.606	0.545	-0.269	0.142	
Domain count	25-29	60 - 64	-0.2352	0.097	-2.417	**0.016**	-0.426	-0.044	
Domain count	25-29	65+	-0.1627	0.084	-1.93	0.054	-0.328	0.003	
URI per hour	25-29	19 - 21	0.1317	0.154	0.853	0.394	-0.171	0.434	
URI per hour	25-29	22 - 24	0.0341	0.09	0.38	0.704	-0.141	0.21	
URI per hour	25-29	30 - 34	-0.0205	0.138	-0.149	0.882	-0.29	0.249	
URI per hour	25-29	35 - 39	-0.4111	0.15	-2.733	**0.006**	-0.706	-0.116	Yes
URI per hour	25-29	40- 44	0.1307	0.091	1.441	0.15	-0.047	0.308	
URI per hour	25-29	45 - 49	-0.1679	0.134	-1.251	0.211	-0.431	0.095	
URI per hour	25-29	50 - 54	0.0676	0.193	0.35	0.726	-0.311	0.446	
URI per hour	25-29	55 - 59	0.0385	0.209	0.185	0.853	-0.37	0.448	
URI per hour	25-29	60 - 64	0.0191	0.152	0.126	0.9	-0.278	0.316	
URI per hour	25-29	65+	0.0906	0.09	1.003	0.316	-0.086	0.268	
Domain per hour	25-29	19 - 21	0.066	0.08	0.824	0.41	-0.091	0.223	
Domain per hour	25-29	22 - 24	0.0951	0.084	1.127	0.26	-0.07	0.261	
Domain per hour	25-29	30 - 34	0.0678	0.087	0.776	0.438	-0.103	0.239	
Domain per hour	25-29	35 - 39	-0.0371	0.089	-0.416	0.678	-0.212	0.138	
Domain per hour	25-29	40- 44	0.0991	0.096	1.037	0.3	-0.088	0.286	
Domain per hour	25-29	45 - 49	-0.0339	0.091	-0.374	0.709	-0.212	0.144	
Domain per hour	25-29	50 - 54	0.1514	0.092	1.641	0.101	-0.029	0.332	
Domain per hour	25-29	55 - 59	-0.0566	0.128	-0.442	0.659	-0.308	0.195	
Domain per hour	25-29	60 - 64	-0.1033	0.102	-1.015	0.31	-0.303	0.096	
Domain per hour	25-29	65+	0.0194	0.088	0.221	0.825	-0.153	0.191	