明治大学総合数理学部

2017 年度

卒　業　研　究

購買履歴データの
特徴量による国籍推定

学位請求者　先端メディアサイエンス学科

田中司

目次

1　はじめに

個人情報保護法は 2005 年 4 月に施行されて以来すでに 10 年余りが経過し，急速な情報通信技術の発展から情報を取り巻く状況は大きく変化した. 個人情報を含むパーソナルデータの取得・収集・分析・流通が社会経済活動及びイノベーションや経済成長における重要な役割を果たすようになった[1]．それに伴って, 2015 年 9 月 3 日に個人情報保護法が改正された．そこで新設された制度は, 特定の個人を識別することができないように個人情報を加工したものを匿名加工情報と定義し，その加工方法を定め，個人情報を第三者に提供する際に匿名加工を施せば個人の許可を必要としない[2]. これはパーソナルデータの円滑な利活用を促進するものである. 一方で，扱うデータによって再識別されやすい項目も異なり, 加工手法の種類は定まっていない．匿名加工手法を採用する上で重要視する匿名化すべき項目を明らかにする必要がある.

そこで，本研究では扱うデータを具体的に定め, そのデータの各項目を分析し, 特徴量を得る. その特徴量によって属性推定手法を提案し, 精度を確かめる.

2　模擬大会

2.1 実用的なデータの収集

アンドロイドアプリケーションを使用してデータを収集した.アプリの制約上一人当たり 19 件の履歴データを収集した.データの概要を表 1 に示す.顧客データは 5 つのデータ項目について,被験者に尋ねて構成した.

表１　使用したデータの概要

個人情報	データ種別	データ件数	データ項目	項目
	マスタデータ	4333 件	顧客 ID	2 桁数値
			伝票 ID	男女
			年月日	1 桁数値
			時分	名称
			製品 ID	名称
			単価	
			数量	名称
	トランザクションデータ	397625 件	顧客 ID	2 桁数地
			伝票 ID	yyyy/mm/dd
			年月日	数値
			時分	名称
			製品 ID	名称
			単価	名称
			数量	名称
			用途	カテゴリ
			使用場所	カテゴリ
			料金	数値

2.2 模擬大会の概要

本研究でこの大会をプチPWSCUPと呼ぶ.プチPWSCUPの概要は2016年4月29日～6月8日の期間に開催した.SUICAデータを扱った[3].5種類の安全性指標及び12種類の有用性指標を定義した.安全性指標を表2,有用性指標を表3とする[4][5].
また,複数の安全性指標及び有用性指標を自動で行うプラットフォームを開発した.プラットフォーム開発の目的は,加工データ,安全性評価及び有用性評価のスクリプトの追加,評価値の算出の円滑化である.そこで,サーバー上に評価プラットフォームを構築した.ユーザーは匿名加工データと指標スクリプトをアップロードし,フォームは元データとアップロードされた匿名加工データと指標スクリプトから評価をして,結果をデータベースに格納する.結果はリアルタイムに更新される.プチ PWSCUP 参加者 4 人は一人につき 10 種類の匿名加工データを作成し提出した.

表２　作成した安全性指標の概要

指標名	特徴
S1_userId_norm.py	id毎の料金ベクトルのノルムが近いものを推定
S2_userId_sum.py	id毎の料金合計値が近いものを推定
S3_userId_use.py	id毎のuseの数のユークリッド距離が近いものを推定
S4_userId_cos.py	id毎のuseの数のコサイン類似度から近いものを推定
S5_userId_use.py	use=1内のレコード単位で比較し，一致したセルの数から推定

表３　作成した安全性指標の概要

指標名	特徴
U1_name.py	id毎の合計値の平均絶対誤差
U2_sma5.py	5日移動平均の平均絶対誤差
U1_mae_fare.R	5日移動平均の平均絶対誤差
U2_mae_record.R	日付ごとのレコード数の差の平均絶対誤差
U3_mae_user.R	ユーザーごとのレコード数の差の平均絶対誤差
U4_AYA.R	レコード番号ごとの合計料金の平均絶対誤差
U5_time.R	回数のレコード数の差の平均絶対誤差

U6_station1.R	駅の登場回数の平均絶対誤差
U7_station2.R	路線の登場回数の平均絶対誤差
U8_AYA2.R	レコード番号ごとのIDの平均絶対誤差
satoshi_U9_use1.R	用途と使用場所の組み合わせの平均絶対誤差
U10_use2.R	用途ごとの合計金額の平均絶対誤差

2.3 模擬大会の順位付け

プチ PWSCUP の順位はプラットフォームから出力された結果をもとに評価を行った. それぞれの指標の評価値に関しては 0 に近い値ほどその指標に対して有効であることを示す. 評価値が 0 に近い順にソートし, 順位をつけた. それぞれの匿名加工データにおいて, 安全性指標と有用性指標の順位をすべて足し合わせ, 指標数で割って平均する. この時, 値が小さいほど各指標に有効な匿名加工データとし, 再び小さい順にソートし, 順位をつける. これが最終的な順位である. 図 1 のようになった. 上から順に 1 位から 8 位まで表している.

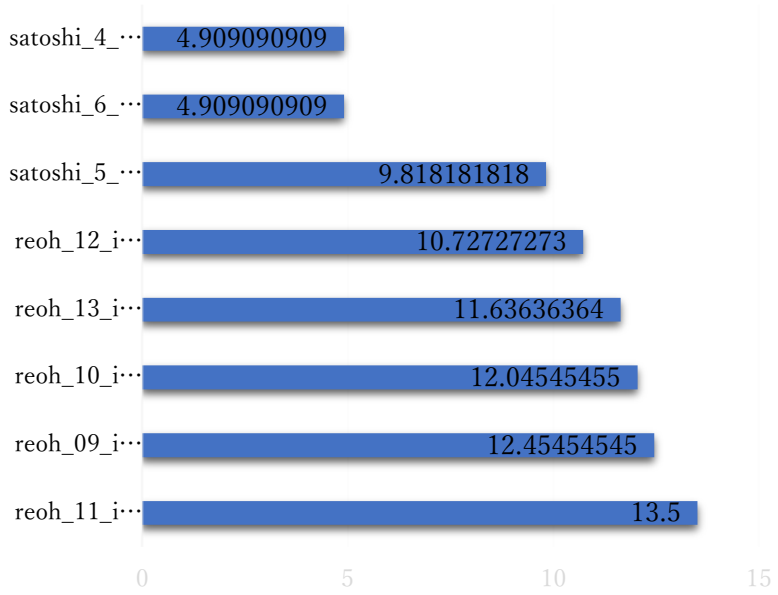


図 1　総合的な評価値

2.4 結果からの分析結果

手法を主に 1. 統一, 2. 平均化, 3. スワップ, 4. 山岡匿名化, 5. 逆順, 6. ランダム化, 7. 偽造タプルの 7 つに分類する． この中でプチ PWSCUP において有効であったのが, 主に 3. スワップと 4. 山岡匿名化であった. 各指標における値を変数として近いもの同士がどれであるか図 2 のように自己組織化を行い可視化した.

この結果からわかることは, それぞれの指標の値が総合的に近い手法すなわち似たような手法が順位的にも近くなっていることを表す.

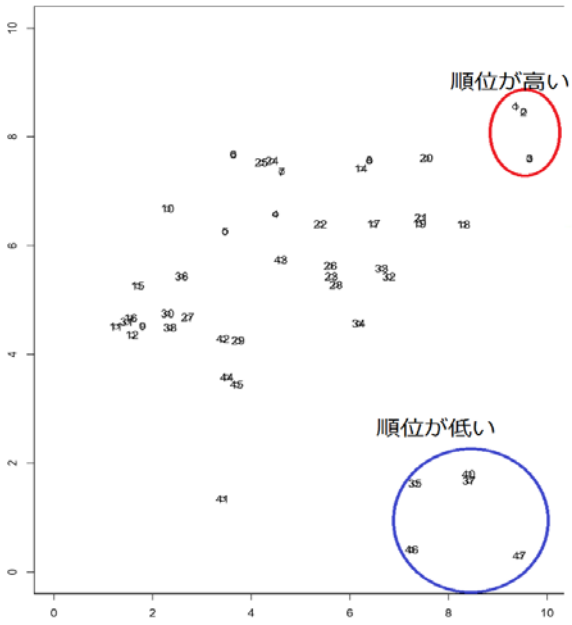


図２　特徴と順位を表す自己組織化マップ

3 購買履歴データの分析

3.1 購買履歴データの概要

本研究は,PWSCUP2016 匿名加工・再識別コンテストで用いられたデータセットを扱う.データセットには顧客データと購買履歴データの 2 種類ある.それぞれのデータの概要について表 4 に示す.また顧客データの例を表 5,購買履歴データの例を表 5 で示す.

表4 使用したデータの概要

	データ種別	データ件数	データ項目	項目
個人情報	顧客データ	4333 件	顧客 ID	5 桁数値
			性別	f m
			生年月日	yyyy/m/d
			国名	名称
	購入履歴データ	397625 件	顧客 ID	5 桁数値
			伝票 ID	6 桁数値
			購入日付	yyyy/m/d
			購入時間	tt:tt
			商品 ID	5 桁数値
			単価	n.nn
			購入数	nn

表5 顧客データの例

顧客 ID	性別	生年月日	国籍
1	f	1995/9/6	Japan
2	m	2000/6/2	China

表6　購買履歴データの例

顧客 ID	伝票 ID	購入日付	購入時間	商品 ID	単価($)	購入数
1	A	1995/9/6	6:00	12	1.2	5
2	B	2000/6/2	12:00	13	2.4	6

表7　総合データの例

伝票 ID	購入日付	購入時間	伝票 ID	単価($)	購入数	性別	生年月日	国籍
A	1995/9/6	6:00	12	1.2	5	f	1995/9/6	Japan
B	2000/9/3	12:00	13	2.4	6	m	2000/6/2	China

3.2 商品の分析

レコード数が最も多い5つの商品の時間による推移を図4に示した.また商品番号と商品名を表8に示した.図3より 22630 と 22629 のグラフが類似している事から商品に関しても類似していると予想できる.実際に表8よる 22630 と 22629 は DOLLY GIRL LUNCH BOX と SPACEBOY LUNCH BOX であり,類似商品である.

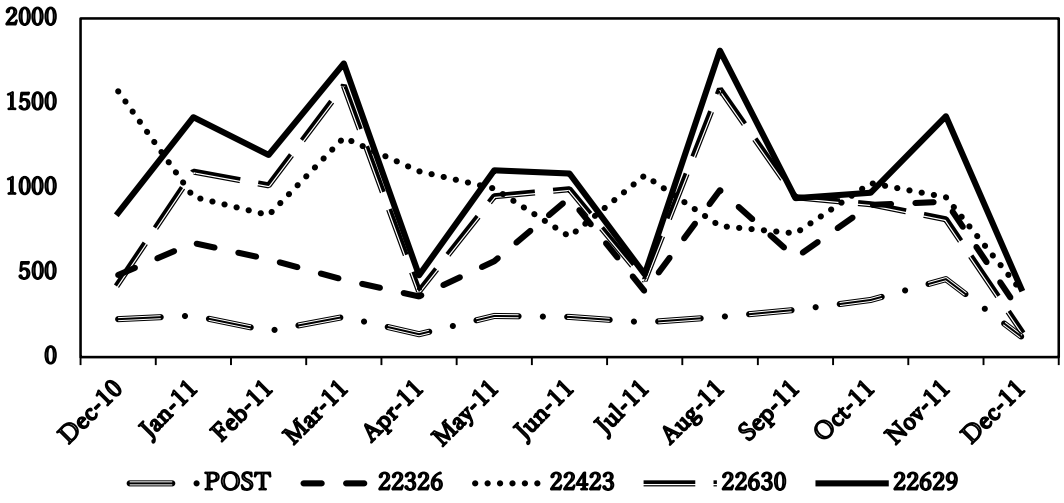


図3 商品の時間による推移

表 8 商品番号と商品名

商品番号	商品名
POST	POSTAGE(送料)
22326	ROUND SNACK BOXES SET OF4 WOODLAND (4 つのウッドランドのラウンドスナックボックス)
22423	REGENCY CAKESTAND 3 TIER (レシピケーキスタンド 3 ティア)
22630	DOLLY GIRL LUNCH BOX (ドリーランチボックス)
22629	SPACEBOY LUNCH BOX (スペースランチボックス)

3.3 国毎の分析方法

国毎に分析するにあたり,購買履歴データには顧客 ID の国籍の項目がない.よって顧客データに含まれる顧客 ID の国籍を購買履歴データに結合させて分析する.そこで購買履歴データと顧客データに対して,顧客 ID を結合時のキーとして結合させる.結合させた後のデータを総合データと呼び,その例を表 7 に示す.総合データを用いて国毎の月別売上個数,時間帯に対する購入頻度,顧客数,合計購入額,顧客あたりの購入額,商品種類数,伝票種類数,顧客あたりの平均伝票数,合計個数,伝票あたりの平均個数,レコード数,平均購入額,平均購入月を求める.顧客あたりの購入額トップ 20 国を図 4 に示す。また,国籍別の月の売上個数の分布を図 5 に示した.

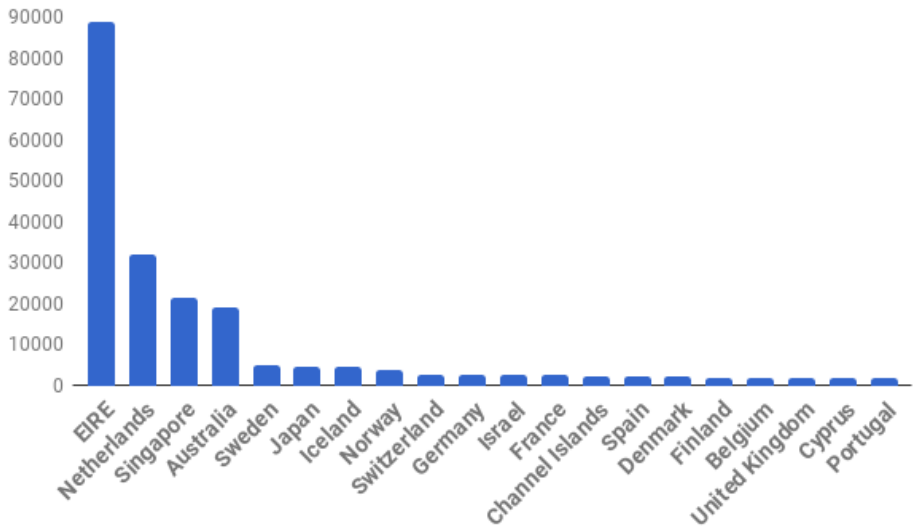
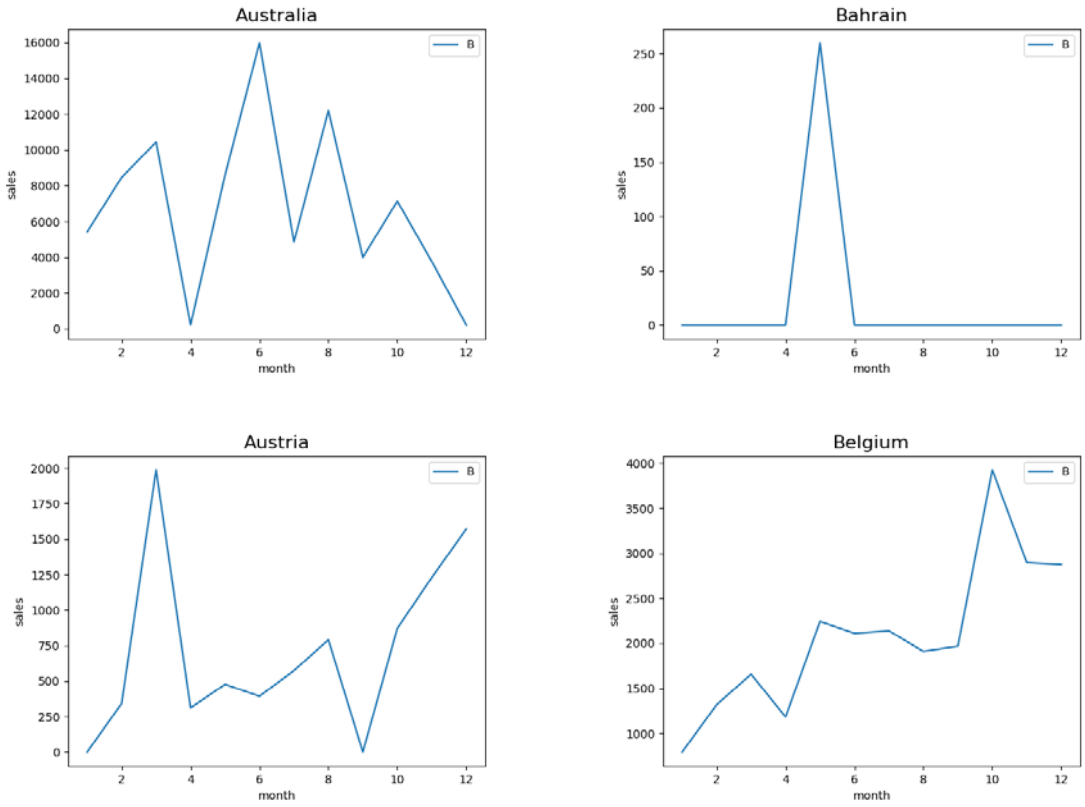
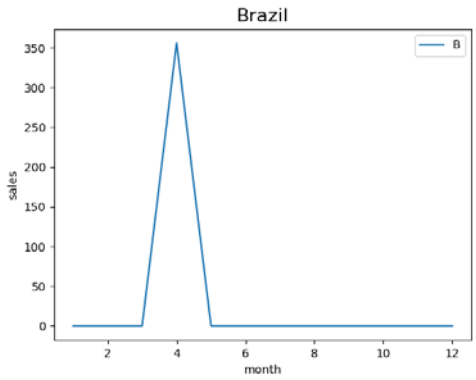
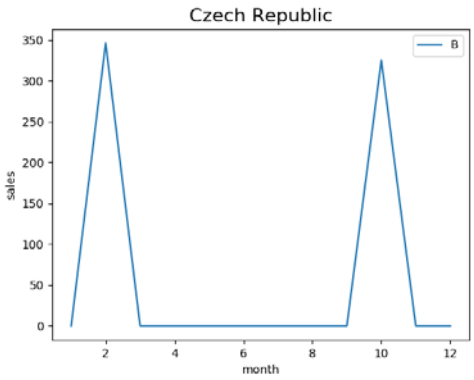
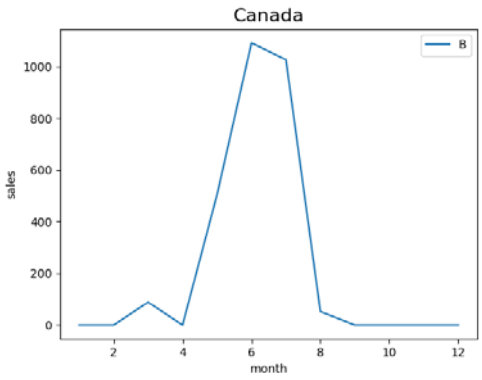


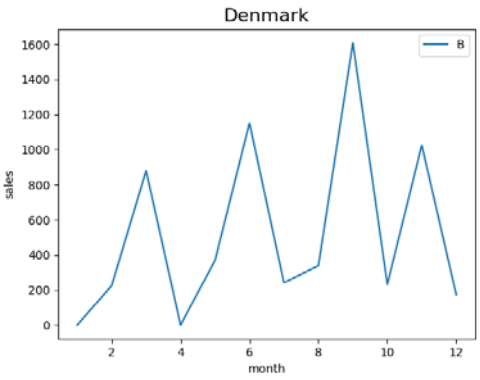
図4　顧客あたりの合計購入額トップ20国

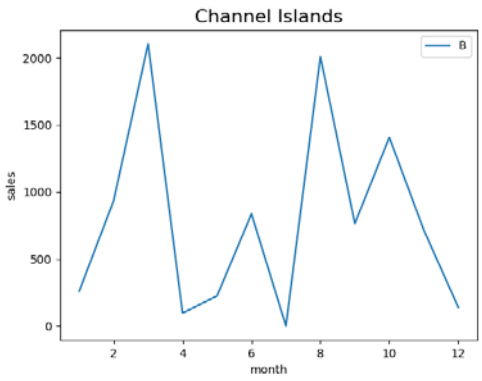


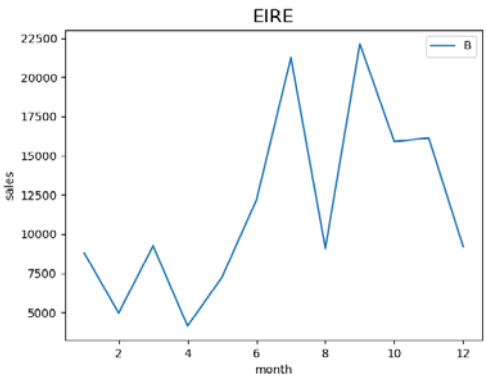

Brazil
B
sales
month

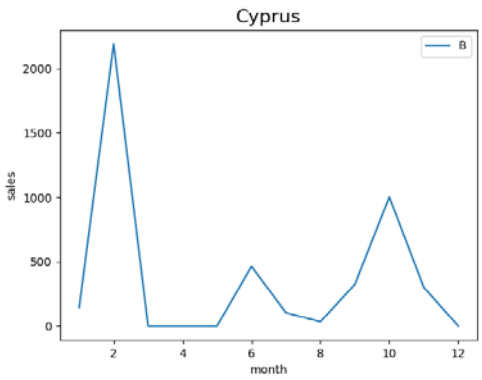

Czech Republic
B
sales
month

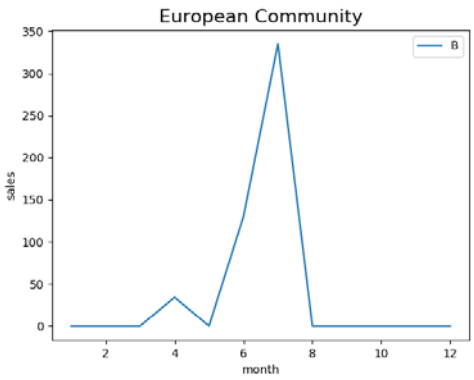

Canada
B
sales
month

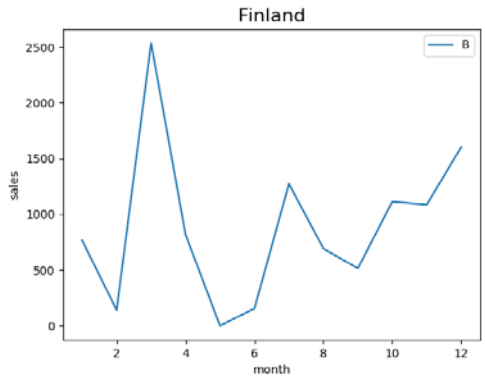

Denmark
B
sales
month

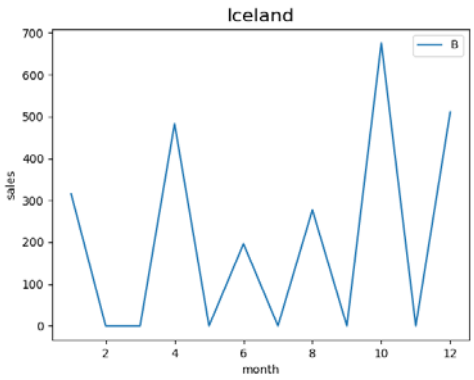

Channel Islands
B
sales
month

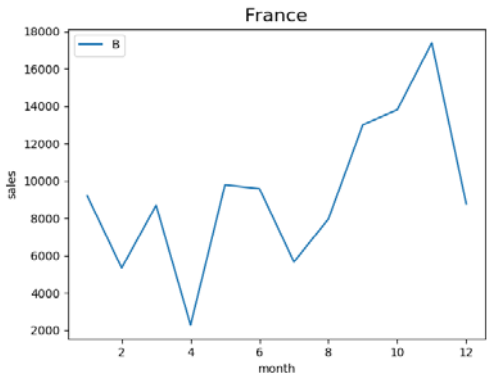

EIRE
B
sales
month

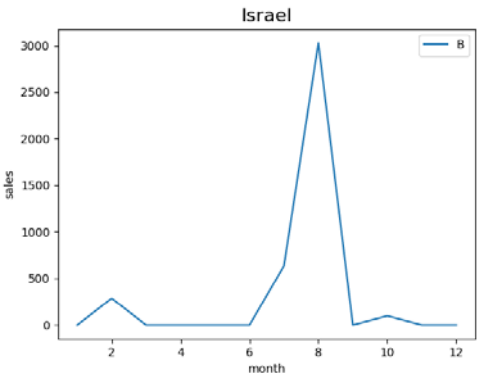

Cyprus
B
sales
month

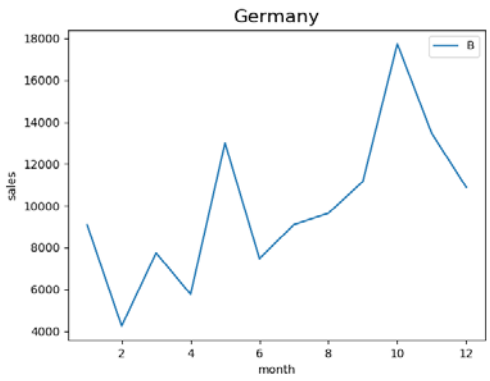

European Community
B
sales
month

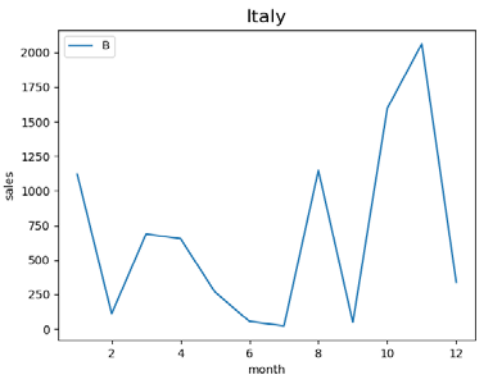

Finland
B
sales
month

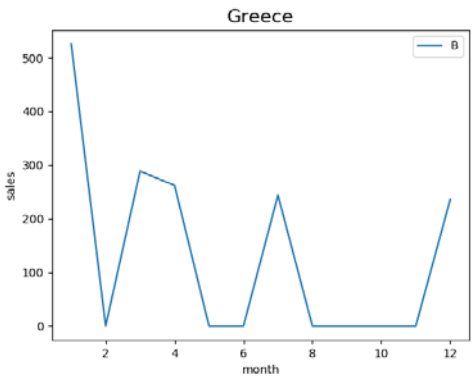

Iceland
B
sales
month

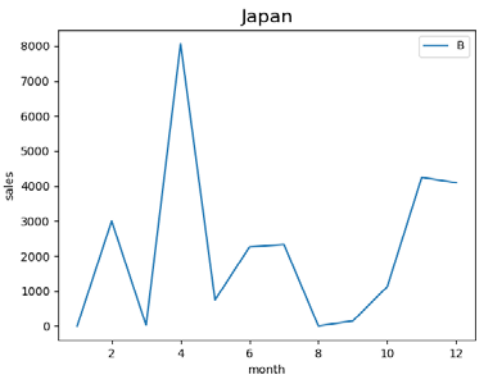

France
B
sales
month

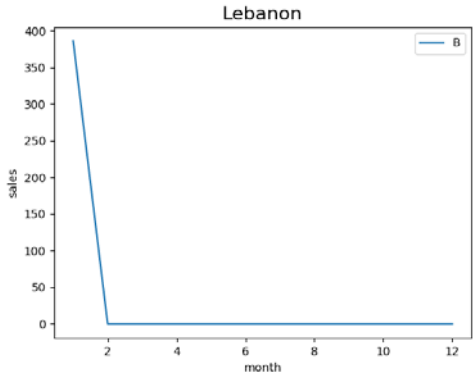

Israel
B
sales
month

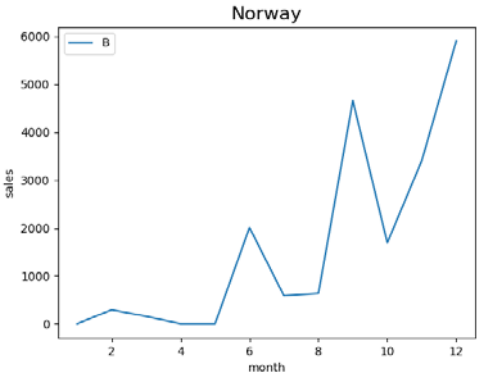

Germany
B
sales
month

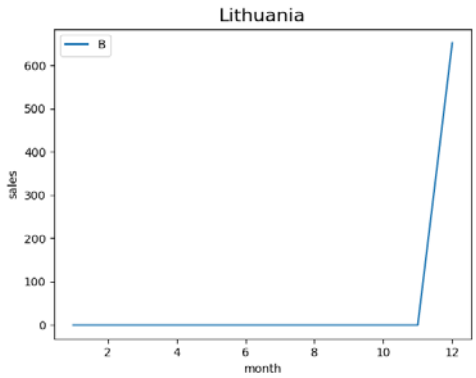

Italy
B
sales
month

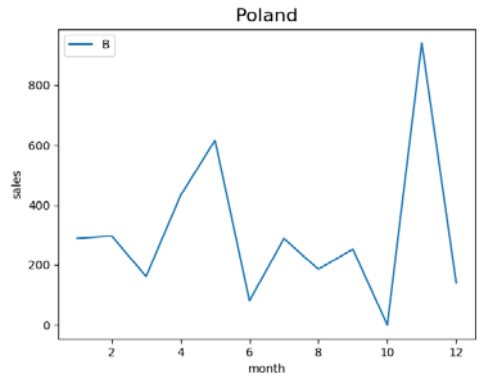

Greece
B
sales
month

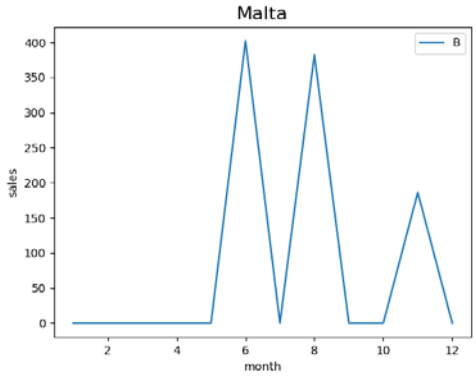

Japan
B
sales
month

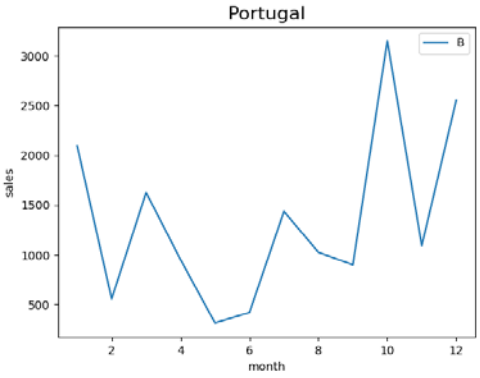

Lebanon
B
sales
month

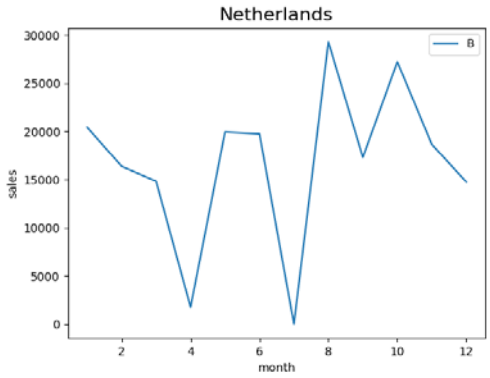

Norway
B
sales
month

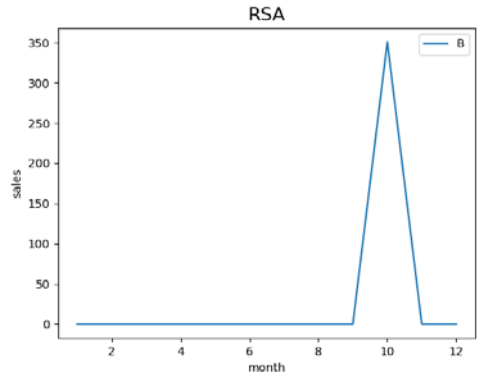

Lithuania
B
sales
month


Poland
B
sales
month


Malta
B
sales
month


Portugal
B
sales
month


Netherlands
B
sales
month


RSA
B
sales
month

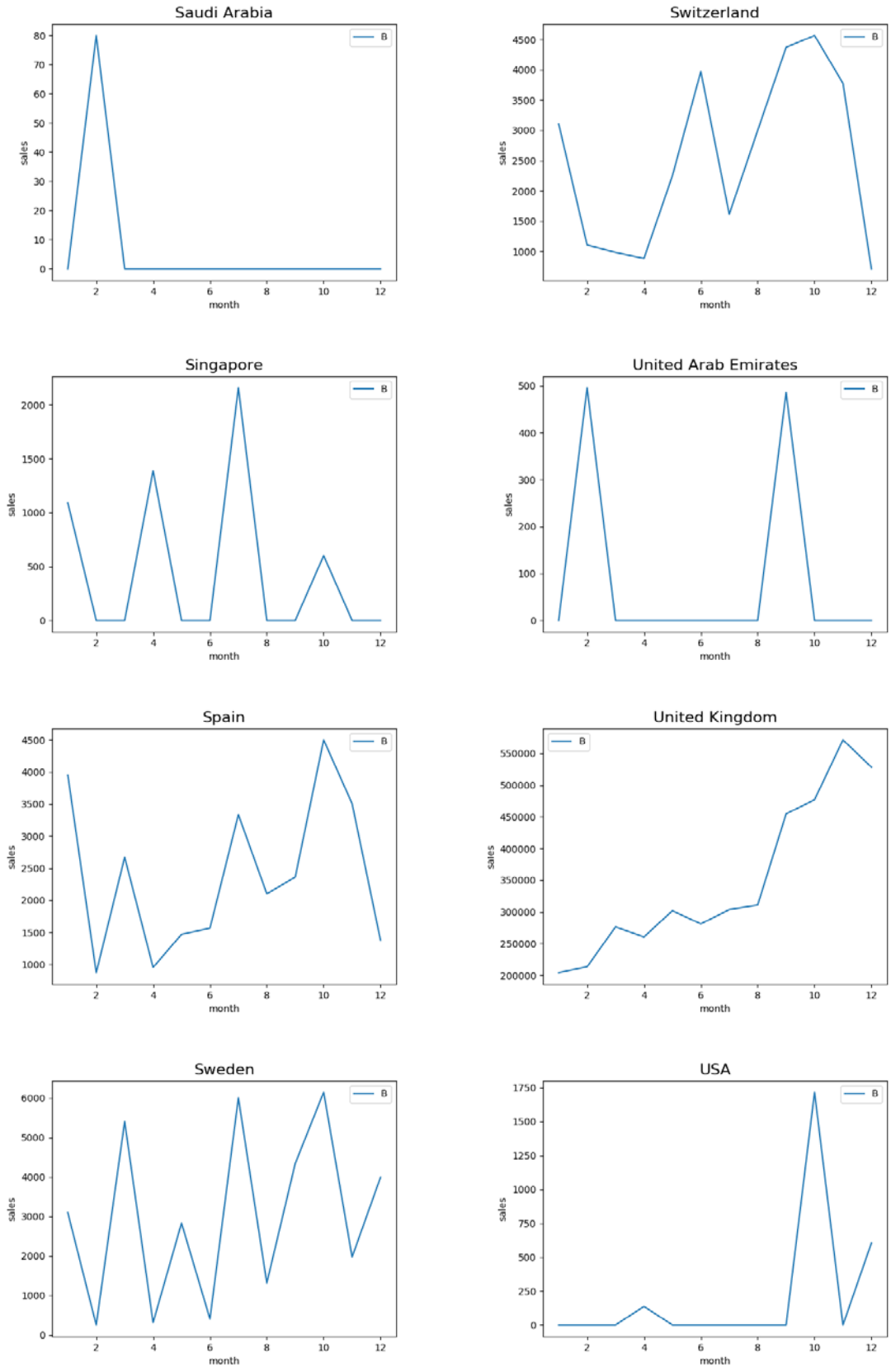
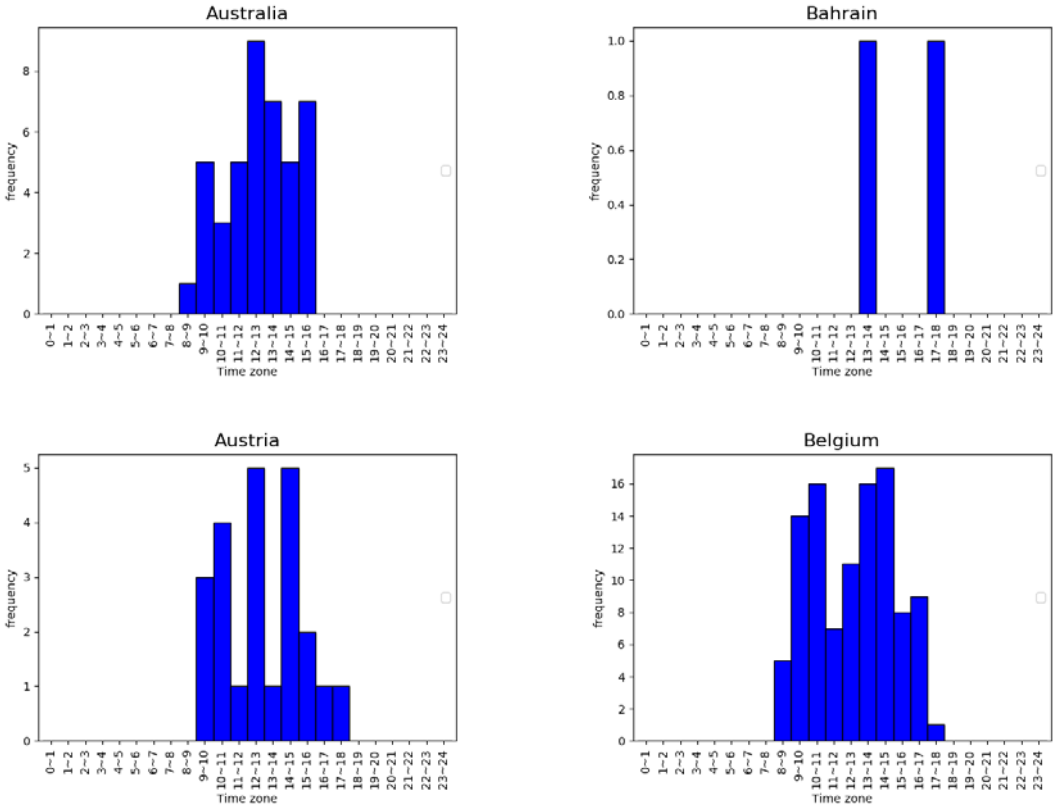
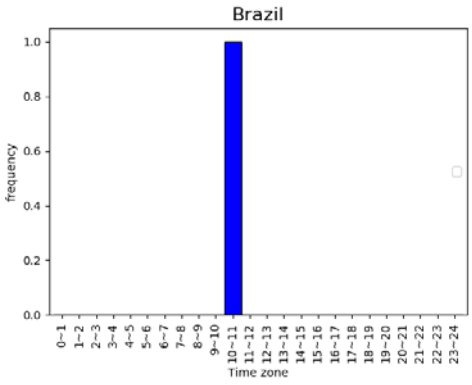


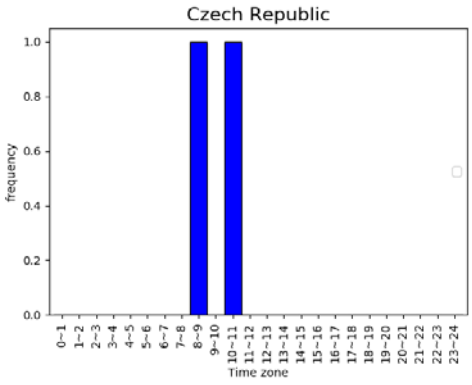
図5 各国の月の売上個数の分布

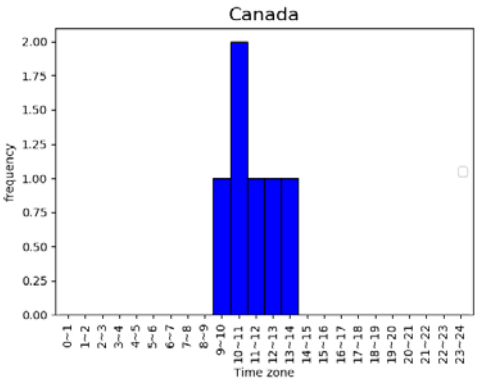
3.4 分析結果

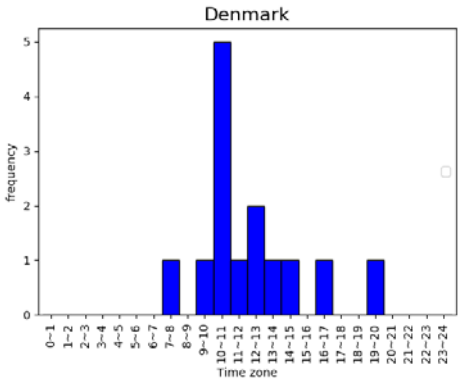
各国の顧客数,合計購入額,取引商品数,伝票数を表 9 に示した.図 3 より顧客あたりの平均購入額の最も高いのがアイルランドである.しかし,顧客数が 3 人で United Kingdom の 3918 人と比べると少ない.よってこの分析結果からアイルランドは顧客合計購入額が多いとは言えず,特異な顧客による影響である可能性がある. 図 5 の国籍毎の月の売上個数は国籍毎にグラフが大きく異なり,特徴量があると言えるが,グラフからすべての国籍の特徴量を決め区別するのは難しい.よってこれについては 4 章で数値を用いて特徴量を表現する.比較また国籍毎の購入回数の時刻帯分布を図 6 に示した.時間帯による購入頻度についてはどの国においても 9 時から 18 時の間に集中していて,国による時差によって生じる購入する時間帯の差は見られなかった.おそらく国籍が異なる顧客でも在住している国はヨーロッパであるから国の時差が生じないと予想できる.

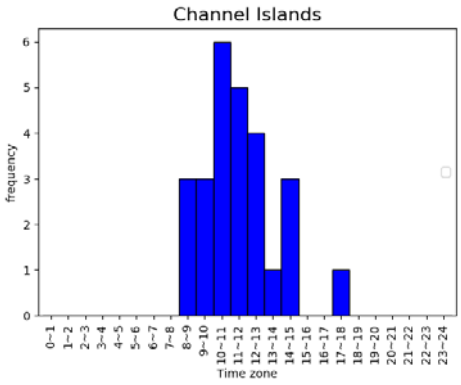


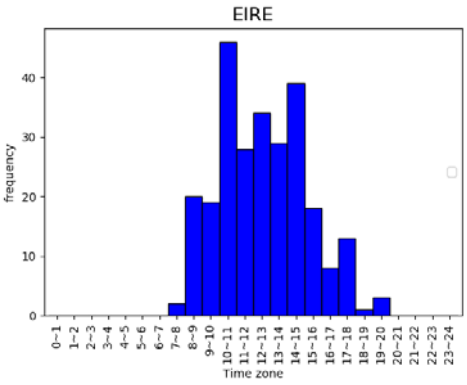

Brazil
frequency
Time zone

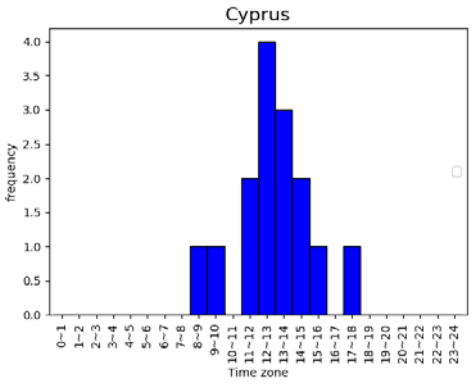

Czech Republic
frequency
Time zone

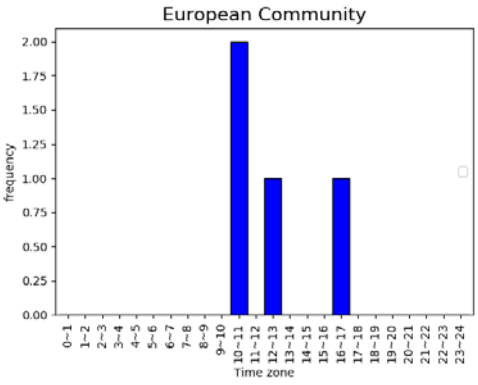

Canada
frequency
Time zone

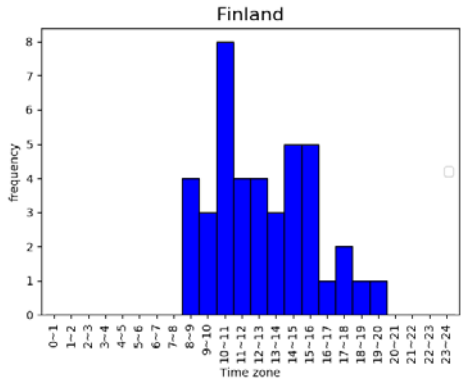

Denmark
frequency
Time zone

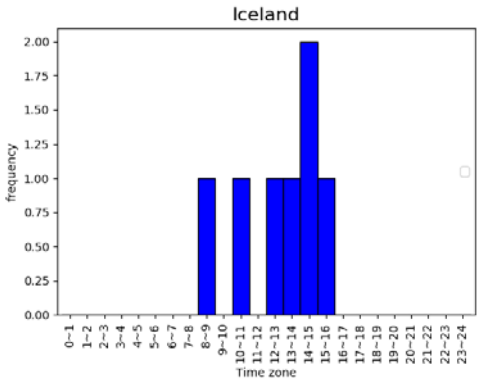

Channel Islands
frequency
Time zone

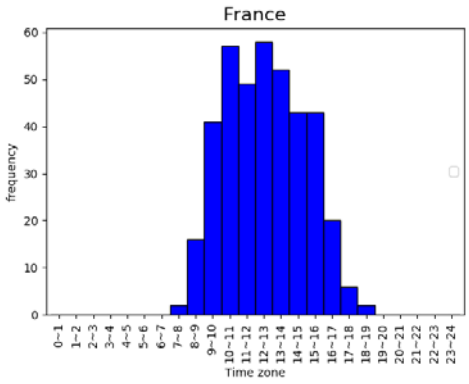

EIRE
frequency
Time zone

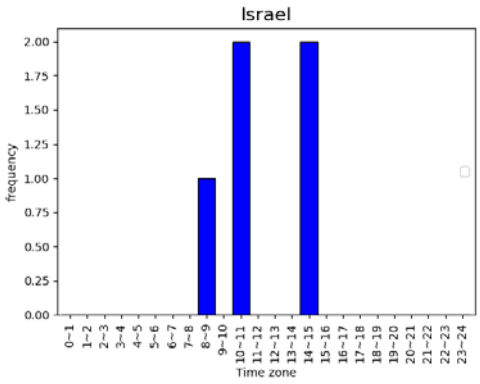

Cyprus
frequency
Time zone

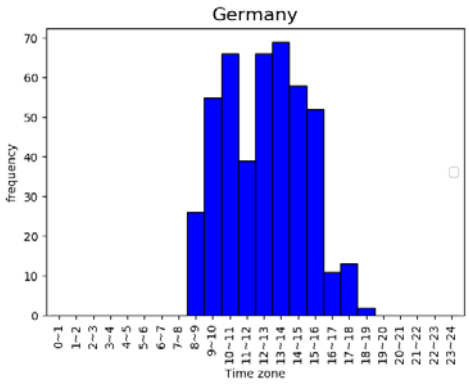

European Community
frequency
Time zone

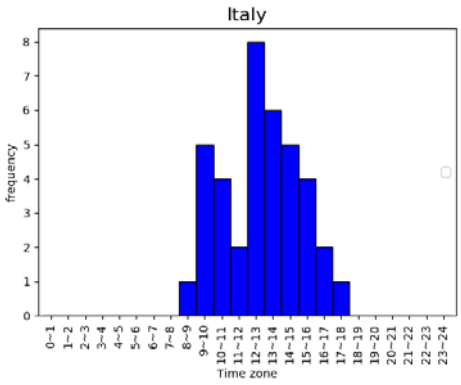

Finland
frequency
Time zone

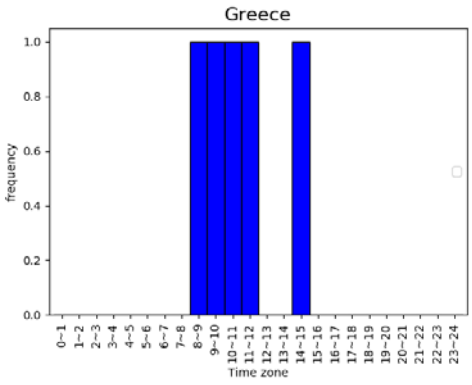

Iceland
frequency
Time zone

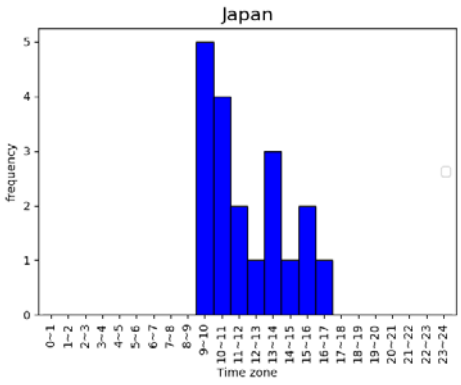

France
frequency
Time zone

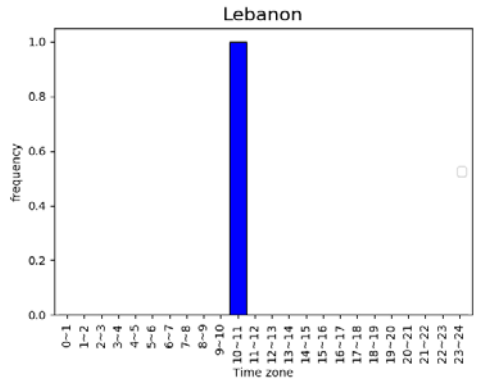

Israel
frequency
Time zone

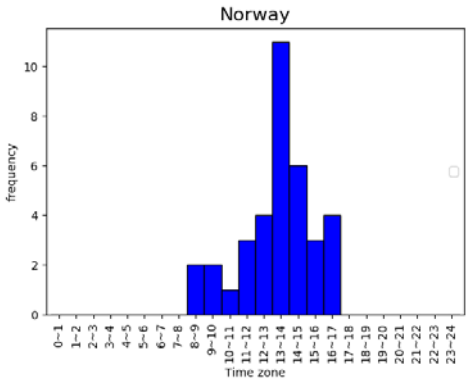

Germany
frequency
Time zone

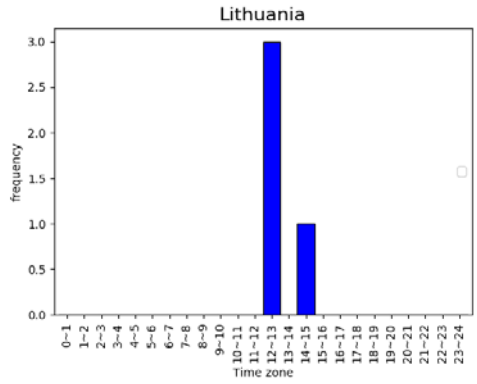

Italy
frequency
Time zone

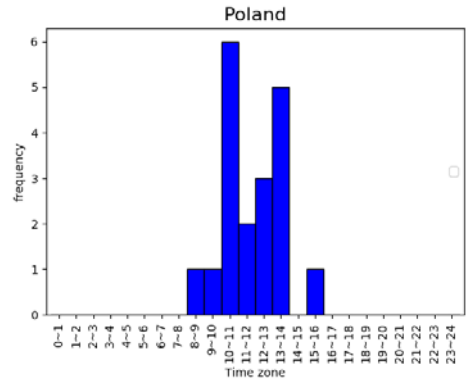

Greece
frequency
Time zone

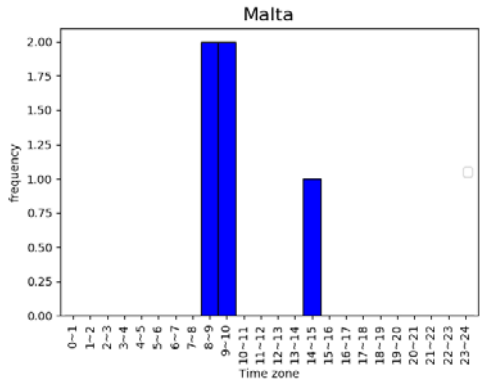

Japan
frequency
Time zone

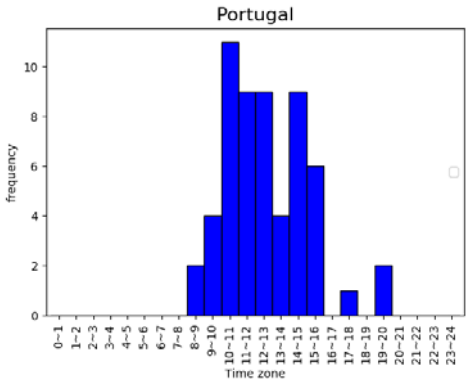

Lebanon
frequency
Time zone

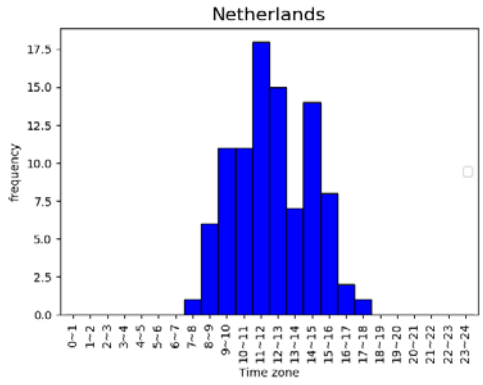

Norway
frequency
Time zone

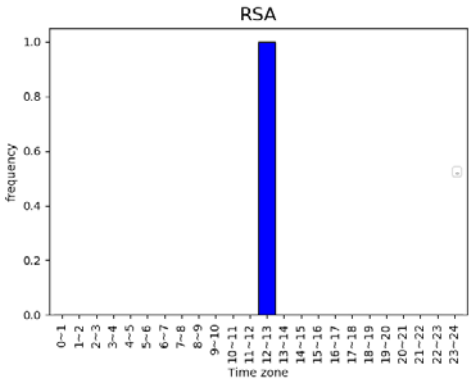

Lithuania
frequency
Time zone


Poland
frequency
Time zone


Malta
frequency
Time zone


Portugal
frequency
Time zone


Netherlands
frequency
Time zone


RSA
frequency
Time zone

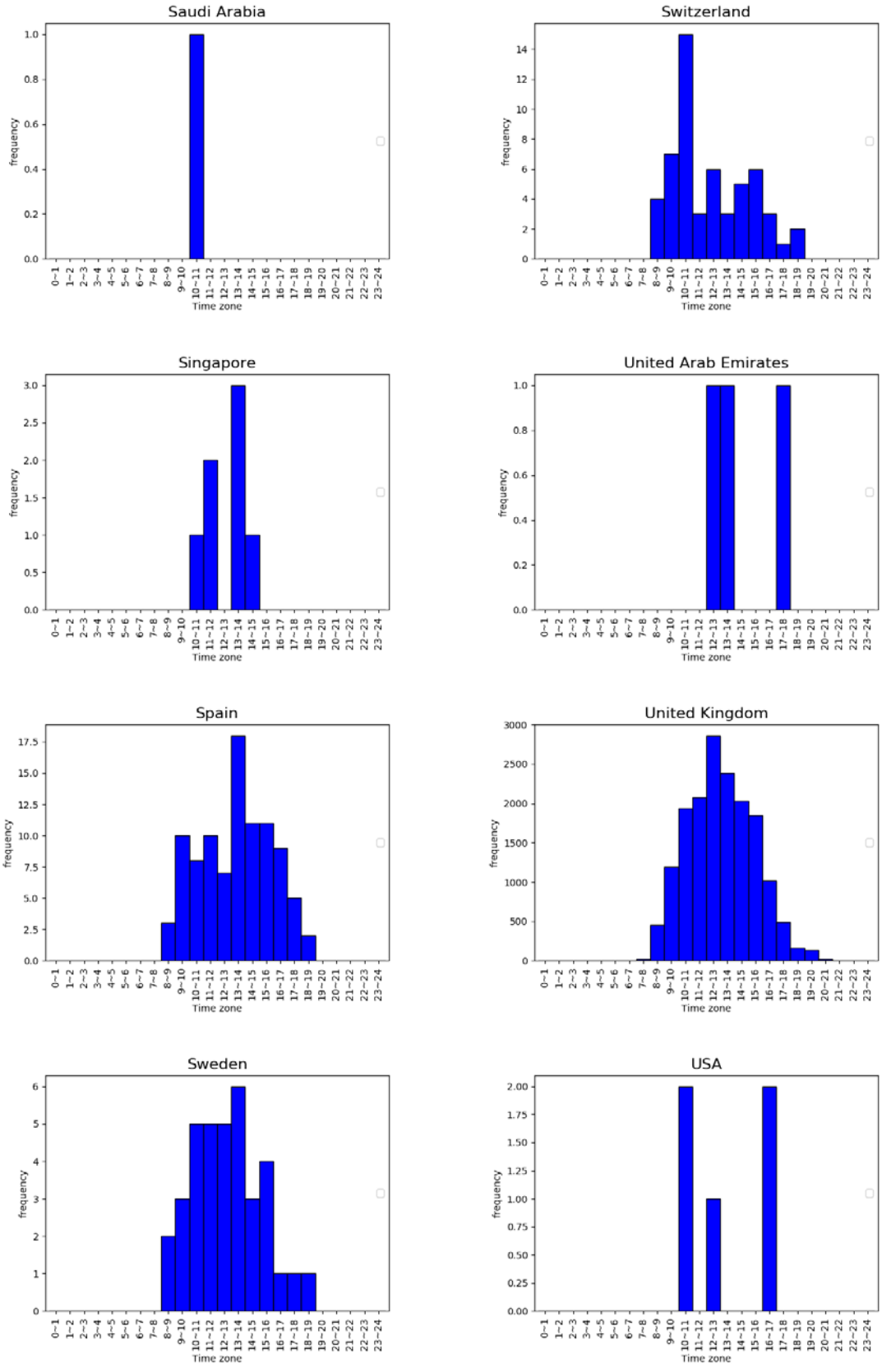


図6　各国においての購入回数の時刻分布

表9 使用したデータの概要

国名	顧客数	合計購入額（$）	取引商品数	伝票数
United Kingdom	3918	7231007.31	3643	16634
France	87	209024.05	1522	389
Belgium	24	44389.61	750	104
Netherlands	9	285446.34	782	94
Germany	94	228867.14	1664	457
Norway	10	36165.44	591	36
EIRE	3	265545.9	1943	260
Switzerland	21	57222.85	951	55
Spain	29	64021.95	1139	94
Poland	6	7334.65	204	19
Portugal	19	33439.89	686	57
Italy	14	17483.24	473	38
Lithuania	1	1661.06	29	4
Japan	8	37416.37	215	19
Iceland	1	4310	103	7
Australia	7	132620.44	543	42
Channel Islands	9	20450.44	430	26
Austria	11	16779.59	472	23
Sweden	8	38378.33	261	36
Finland	12	22546.08	458	41
Cyprus	6	10697.15	418	15
Greece	4	4760.52	138	5
Singapore	1	21279.29	178	7
Lebanon	1	1693.88	45	1
United Arab Emirates	2	1902.28	68	3
Israel	3	7221.69	219	5
Denmark	7	14751.52	204	14
Saudi Arabia	1	145.92	9	1

Czech Republic	1	826.74	25	2
Canada	4	3666.38	147	6
Brazil	1	1143.6	32	1
USA	4	3580.39	163	5
European Community	1	1300.25	50	4
Bahrain	2	548.4	16	2
Malta	2	2725.59	99	5
RSA	1	1002.31	57	1

4 属性推定

4.1 分析による特徴量

月別売上個数についての国の特徴量を顧客の購買履歴から国籍の推定をするために各月の売上が 0 であるかどうかに着目する。その月の売り上げ個数が 0 の場合は特徴がありとして 1,それ以外は 0 とする.それらを 2 進数 12 桁で表し,10 進数に変換した値を国による購入特徴 ID と定義する.表 5 に国と購入特徴 ID の例を示す.例えば USA の場合,売上個数が 0 である月は 1,2,3,5,6,7,8,9,11 の月であり,2 進数で 111011111010 であり,10 進数に変換すると 1527 である.これが USA の購入特徴 ID である.

4.2 国籍推定手法

4.1節で定義した購入特徴IDを利用してそれぞれの国籍に対して購入特徴IDをプログラム言語Pythonでプログラムを作って国籍推定を試みる.国籍数は36種類あり,購入特徴IDが一意である国が24種類あった.顧客IDに対しても月別売上個数から購入特徴IDを割り出し,国籍と顧客IDの購入特徴IDが一致すればその国籍であると予想できる.表10に国籍と購入特徴IDを示す.

表10 国籍と購入特徴IDと2進数

国籍	購入特徴ID	2進数
Norway	25	11001
Poland	512	1000000000
Lithuania	2047	11111111111
Japan	129	10000001
Iceland	1366	10101010110
Channel Islands	64	1000000
Austria	257	100000001
Finland	16	10000
Cyprus	2076	100000011100
Greece	1970	11110110010
Singapore	3510	110110110110

Lebanon	4094	111111111110
United Arab Emirates	3837	111011111101
Israel	3389	110100111101
Denmark	9	1001
Saudi Arabia	4093	111111111101
Czech Republic	3581	110111111101
Canada	3851	111100001011
Brazil	4087	111111110111
USA	1527	10111110111
European Community	3991	111110010111
Bahrain	4079	111111101111
Malta	2911	101101011111
RSA	3583	110111111111

4.3 国籍推定の精度

実際のPWSCUP2016の本戦で提出された加工済み購買履歴データを用いて国籍推定手法の精度を確かめる.まず加工済み購買履歴データのすべての顧客IDの月別売上個数を求め,購入特徴IDを算出する.国籍の購入特徴IDと一致した顧客IDは90種類あった.実際の国籍と比較すると90種類全て再識別されなかった.精度は0%である.国籍の購入履歴の特徴が加工済み購買履歴データの顧客IDの特徴と一致する確率は極めて低いことがわかった.

5 おわりに

研究室 31 人のデータは小規模でありながら充分な研究ができた. その結果, 山岡匿名化と呼ばれる指標の穴をつくような手法があり, これを見抜く事も今後の課題である. 加工購買履歴データにおいて顧客 ID の国籍を推定する事は極めて難しかった. 国籍は顧客の集合の売上個数であるので必ずしも国の特徴とその国籍の顧客の特性は一致しない. 本研究で用いたデータは国による顧客数に偏りがあり, 全ての国に対してその国独自の特徴であるといえる顧客人数ではなかったことが原因と考える. 国籍の候補を絞り出し、他の複数の国籍の特徴を組み合わせて精度を向上することが今後の課題である.

謝辞

本研究においてご指導を頂いた菊池浩明教授に感謝致します．また, 研究にお付き合い頂いた本研究室先輩方の伊藤聡志氏, 原田玲央, 岡本健太郎氏, 後輩の小林祐貴及び菊池研究室の皆様に感謝致します.

参考文献

[1] 個人情報保護委員会事務局,パーソナルデータの利活用促進と消費者の信頼性確保に向けて,[online] https://www.ppc.go.jp/files/pdf/report_office.pdf(参照 2017 年 5 月)

[2] 経済産業省,事業者が匿名加工情報の具体的な作成方法を検討するにあたっての参考資料,[online] http://www.meti.go.jp/press/2016/08/20160808002/20160808002-1.pdf(参照 2016 年 9 月)

[3] 菊池浩明, 小栗 秀暢, 野島 良, 濱田 浩気, 村上 隆夫, 山岡 裕司, 山口 高康, 渡辺知恵美, “PWSCUP:履歴データを安全に加工せよ”, CSS2016, pp.271-278, 2016.

[4] 伊藤聡志,乗降履歴データの有用性評価指標と匿名加工,明治大学総合数理学部 2016 年度卒業論文

[5] 原田玲央,商品の特徴による再識別リスクとクラスタリングを用いた購買履歴データ匿名加工手法の提案,明治大学総合数理学部 2016 年度卒業論文