GLM for partially pooled categorical pre

GLM

for

par-

tially

pooled

cat-

e-

gor-

i-

cal

pre-

dic-

tors

with

a

case

study

in

biose-

cu-

rity

Christopher M. Baker[1,2,3], Hov

1

[3] Centr

[4] D

GLM for partially pooled categorical pre

SUMMARY:

Na-

tional

govern-

ments

use

bor-

der

in-

for-

ma-

tion

to

ef-

fi-

ciently

man-

age

the

biose-

cu-

rity

risk

pre-

sented

GLM for partially pooled categorical pre

by

travel

and

com-

merce.

In

the

Aus-

tralian

bor-

der

biose-

cu-

rity

sys-

tem,

GLM for partially pooled categorical pre

data

about

cargo

en-

tries

are

col-

lected

from

records

of

di-

rec-

tions:

that

is,

the

records

of

ac-

tions

taken

GLM for partially pooled categorical pre

by

the

biose-

cu-

rity

reg-

u-

la-

tor.

An

en-

try

is

a

col-

lec-

tion

of

im-

port

lines

where

each

line

GLM for partially pooled categorical pre

is

a

sin-

gle

type

of

item

or

com-

mod-

ity.

Anal-

y-

sis

GLM for partially pooled categorical pre

is

sim-

ple

when

the

data

are

recorded

in

line

mode:

the

di-

rec-

tions

are

recorded

in-

di-

vid-

u-

ally

for

each

line.

The

chal-

lenge

comes

when

GLM for partially pooled categorical pre

data

are

recorded

in

con-

tainer

mode,

be-

cause

the

same

di-

rec-

tion

is

recorded

against

each

line

in

the

en-

try,

mean-

ing

that

we

don?t

know

which

line(s)

within

the

en-

try

are

non-

compliant.

We

de-

velop

a

sta-

tis-

ti-

cal

model

to

use

con-

tainer

mode

GLM for partially pooled categorical pre

data

to

GLM for partially pooled categorical pre

help

in-

form

biose-

cu-

rity

risk

of

items.

We

use

asymp-

totic

anal-

y-

sis

to

es-

ti-

mate

the

value

of

con-

tainer

mode

GLM for partially pooled categorical pre

data

com-

pared

to

line

mode

data,

do

a

sim-

u-

la-

tion

study

to

ver-

ify

that

we

can

ac-

cu-

rately

es-

ti-

mate

pa-

ram-

e-

ters

in

a

large

dataset,

and

we

ap-

ply

our

methods

to

a

real

dataset,

for

which

im-

por-

tant

in-

for-

ma-

tion

about

the

risk

of

non-

compliance

is

re-

covered

us-

ing

the

new

model.

KEY

WORDS:

GLM for partially pooled categorical pre

Bayesian

in-

fer-

ence;

Risk

anal-

y-

sis;

Cargo.

Biometrics

000,

??–
??

DOI:

000
000

0000

This paper has been submitted

1. Introduction

Invasive

species

pose

a

mul-

ti-

faceted

threat

to

SO-

ci-

ety,

lead-

ing

to

re-

duc-

tions

in

agri-

cul-

tural

pro-

duc-

tiv-

ity,

as

well

as

damages

to

the

en-

vironment,

human

health,

and

the

economy

(?).

Con-

sid-

er-

able

ef-

fort

is

de-

voted

to

man-

ag-

ing

in-

va-

sive

species

(?),

in

ei-

ther

erad-

i-

cat-

ing

them

(????),

or

suppress-

ing

their

num-

bers

to

re-

duce

damages

(???).

The

costs

as-

SO-

ci-

ated

with

man-

ag-

ing

in-

va-

sive

species

pro-

vides

govern-

ments

with

an

in-

cen-

tive

to

man-

age

biose-

cu-

rity

risks

at

na-

tional

bor-

ders

to

pre-

vent

the

es-

tablish-

ment

of

the

new

species.

Given

the

mas-

sive

scale

of

global

trade,

biose-

cu-

rity

reg-

u-

la-

tors

need

to

be

able

to

al-

lo-

cate

their

re-

sources

ef-

fi-

ciently,

but

to

do

SO

they

must

un-

der-

stand

the

risks

as-

so-

ci-

ated

with

var-

i-

ous

en-

tries.

At

the

bor-

der,

one

of

the

most

ef-

fec-

tive

ways

of

achiev-

ing

this

is

by

us-

ing

the

out-

come

of

pre-

viously

con-

ducted

in-

spec-

tions

as

in-

tel-

li-

gence

that

in-

forms

fu-

ture

op-

er-

a-

tions.

This

al-

lows

reg-

u-

la-

tors

to

iden-

tify

high

risk

com-

modi-

ties

and

im-

porters,

and

mod-

ify

their

in-

spec-

tion

targets

and

policies

ac-

cord-

ingly.

However,

gain-

ing

in-

sight

from

bor-

der

in-

spec-

tions

is

a

mas-

sive

lo-

gis-

ti-

cal

chal-

lenge,

as

data

must

be

con-

sistently

recorded

in

a

for-

mat

that

makes

this

anal-

y-

sis

pos-

si-

ble.

Putting

in-

frastruc-

ture

into

place

to

col-

lect

this

data,

and

col-

lect-

ing

it

ac-

cu-

rately

can

be

ex-

pen-

sive

and

chal-

leng-

ing.

Of-

ten,

shortcuts

may

be

taken

that

ren-

der

the

data

less

valu-

able

in

analysing

patterns

of

risk.

In

Aus-

tralia

sys-

tems

have

been

put

in

place

to

cap-

ture

biose-

cu-

rity

data

since

the

early

90's;

these

data

are

ex-

tracted

from

the

di-

rec-

tions

ap-

plied

to

the

cargo

in

ques-

tion.

Di-

rec-

tions

are

used

to

con-

trol

the

move-

ment

and

di-

rect

the

as-

sess-

ment

and

man-

age-

ment

of

goods

sub-

ject

to

biose-

cu-

rity

con-

trol,

and

the

GLM for partially pooled categorical pre

modes

dis-

cussed

in

this

pa-

per

cor-

re-

spond

to

how

those

di-

rec-

tions

are

carried

out:

at

the

line

level,

or

at

the

en-

try

level.

As

such,

in

line

mode,

the

de-

tails

of

which

items

within

an

en-

try

are

in-

spected

and

found

to

be

com-

pli-

ant

or

non-

compliant

are

fully

recorded.

How-

ever,

in

con-

tainer

mode,

the

re-

sults

of

an

in-

spec-

tion

are

ap-

plied

to

all

items

within

an

en-

try.

Con-

tainer

mode

was

in-

troduced

SO

that

con-

tain-ers

could

be

re-

leased

piecemeal

in

or-

der

to

minimise

the

bottlenecks

cre-

ated

in

ports

when

lines

that

com-

prise

many

con-

tain-ers

are

held

un-

til

all

of

the

line

has

been

cleared.

Con-

tainer

mode

makes

it

much

quicker

for

bor-

der

staff

to

man-

age

en-

tries

with

a

large

num-

ber

of

items,

but

means

that

when

the

data

are

anal-

ysed,

en-

tries

in

con-

tainer

mode

are

censored

in

these

cases,

it

is

un-

known

which

items

were

in-

spected,

and

which

of

those

were

found

to

be

com-

pli-

ant

or

non-

compliant.

This

makes

anal-

y-

sis

of

the

data

to

iden-

tify

trends

in

biose-

cu-

rity

risk

chal-

leng-

ing.

Data

that

have

been

col-

lected

in

con-

tainer

mode

are

closely

re-

lated

to

the

data

col-

lected

un-

der

pooled

GLM for partially pooled categorical pre

test-

ing,

which

is

of-

ten

used

for

dis-

ease

surveillance.

Within

the

pooled

test-

ing

lit-

er-

a-

ture

there

are

two

main

branches:

one

aims

to

iden-

tify

pos-

i-

tives

within

a

pool,

while

the

other

seeks

to

use

pooled

data

to

es-

ti-

mate

quan-

ti-

ties

about

the

pop-

u-

la-

tion.

It

is

the

lat-

ter

es-

ti-

mat-

ing

quan-

ti-

ties

that

we

are

in-

ter-

ested

in.

The

funda-

men-

tal

prob-

lem

is

es-

ti-

mat-

ing

a

preva-

lence,

$p,$

within

a

pop-

u-

la-

tion,

when

only

pooled

data

are

avail-

able

(?).

More

re-

cent

work

has

fo-

cused

on

im-

prov-

ing

es-

ti-

mates

by

re-

duc-

ing

bias,

ei-

ther

through

al-

ter-

ing

the

sam-

pling

strategy

(??)

or

by

in-

cor-

porat-

ing

bias

cor-

rec-

tion

into

mod-

els

(??).

There

have

also

been

ex-

ten-

sions

of

the

prob-

lem

where

p

is

not

a

con-

stant,

but

it

is

es-

ti-

mated

us-

ing

linear

re-

gres-

sion

us-

ing

only

the

pooled

data

(????).

These

pa-

pers

have

made

signif-

i-

cant

progress

in

fit-

ting

in-

creasingly

com-

plex

mod-

els,

but

do

not

fo-

cus

on

the

im-

pacts

of

dif-

fer-

ent

types

of

pool-

ing

on

the

pre-

ci-

sion

of

model

es-

ti-

mates.

In

practice,

how-

ever,

pooled

test-

ing

is

man-

i-

festly

dif-

fer-

ent

from

the

biose-

cu-

rity

sce-

nario.

Pooled

test-

ing

ex-

ists

GLM for partially pooled categorical pre

by

de-

sign:

as

a

way

to

gather

in-

for-

ma-

tion

about

a

pop-

u-

la-

tion

while

re-

duc-

ing

test-

ing.

In

biose-

cu-

rity,

in-

spec-

tion

is

ap-

plied

to

ev-

ery

in-

di-

vidual

line,

and

the

re-

sults

are

only

pooled

at

the

point

of

data

cap-

ture

as

a

side

ef-

fect

of

the

mode

se-

lec-

tion

of

the

en-

try.

Hence,

we

are

in-

ter-

ested

in

how

much

in-

for-

ma-

tion

we

are

los-

ing

due

to

ag-

gregat-

ing

re-

sults

as

con-

tainer

mode.

Our

anal-

y-

sis

of-

fers

reg-

u-

la-

tors

the

op-

por-

tunity

to

as-

sess

the

risks

of

the

con-

tinued

use

of

con-

tainer

mode,

and

to

weigh

them

against

its

op-

er-

a-

tional

advan-

tages.

In

this

pa-

per

we

in-

ves-

ti-

gate

the

ef-

fect

of

con-

tainer

mode

data

col-

lec-

tion

upon

our

abil-

ity

to

es-

ti-

mate

the

biose-

cu-

rity

risk

of

items.

We

start

with

an

asymp-

totic

anal-

y-

sis,

where

we

cal-

cu-

late

the

pre-

ci-

sion

of

es-

ti-

mates

and

de-

ter-

mine

the

im-

pli-

ca-

tions

of

mix-

ing

dif-

fer-

ent

item

types

in

con-

tainer

mode.

We

then

de-

velop

a

sim-

u-

la-

tion

ex-

per-

i-

ment

that

al-

lows

us

to

un-

der-

stand

how

larger

en-

tries

and

more

item

types

af-

fect

the

pre-

ci-

sion

of

our

es-

ti-

mates.

Fi-

nally,

we

anal-

yse

biose-

cu-

rity

data

pro-

vided

GLM for partially pooled categorical pre

by

the

Aus-

tralian

De-

part-

ment

of

Agri-

cul-

ture,

Fisheries

and

Forestry

(DAFF)

to

iden-

tify

the

real-

world

dif-

fer-

ences

be-

tween

us-

ing

the

line-

only

data

and

in-

clud-

ing

the

con-

tainer

mode

data.

2. Model overview

2.1 *Data*

To make

our

language

about

the

data

more

pre-

cise,

we

will

ex-

plicitly

de-

fine

what

we

mean

by

en-

tries,

lines

and

di-

rec-

tions.

En-

tries

are

a

col-

lec-

tion

of

lines,

and

a

line

is

a

group

of

the

same

type

of

item

or

com-

mod-

ity

be-

ing

im-

ported.

When

cargo

en-

ters

the

coun-

try,

each

line

is

given

di-

GLM for partially pooled categorical pre

rec-

tions.

These

di-

rec-

tions

de-

tail

all

of

the

ac-

tiv-

i-

ties

un-

der-

taken

GLM for partially pooled categorical pre

by

the

biose-

cu-

rity

reg-

u-

la-

tor

to

man-

age

the

biose-

cu-

rity

risk

of

each

line,

and

also

the

out-

come

of

those

ac-

tiv-

i-

ties.

To

iso-

late

the

un-

certainty

that

arises

from

use

of

con-

tainer

mode,

we

fo-

cus

on

one

as-

pect

of

the

di-

rec-

tions

recorded

against

im-

port

lines:

whether

they

were

deemed

com-

pli-

ant

(within

biose-

cu-

rity

reg-

u-

la-

tions)

or

not.

Line

and

con-

tainer

mode

are

the

two

ways

that

records

are

kept

of

ac-

tions

made

GLM for partially pooled categorical pre

by

the

biose-

cu-

rity

reg-

u-

la-

tor.

In

line

mode,

the

di-

rec-

tions

as-

signed

to

each

line

are

recorded

along

with

the

out-

come

for

that

line.

In

con-

tainer

mode,

di-

rec-

tions

are

only

recorded

per

en-

try.

This

means

that

in

con-

tainer

mode,

if

any

line

in

that

en-

try

has

an

in-

spec-

tion,

and

non-

compliance

is

found,

then

ev-

ery

line

in

that

en-

try

is

recorded

to

be

non-

compliant.

If

all

of

the

lines

in

the

en-

try

are

com-

pli-

ant,

then

they

are

all

marked

com-

pli-

ant,

SO

in

this

case

line

and

con-

tainer

mode

are

equiv-

a-

lent.

2.2 *Modelling*

Throughout

this

pa-

per,

we

fo-

CUS

on

es-

ti-

mat-

ing

the

prob-

a-

bil-

ity

that

a

line

is

non-

compliant

us-

ing

in-

for-

ma-

tion

in-

clud-

ing

the

type

of

item,

coun-

try

of

origin

and

whether

it

has

com-

plete

doc-

u-

men-

ta-

tion.

The

full

model

for

the

prob-

a-

bil-

ity

that

a

line

is

non-

compliant,

$p_{ijk\ell}$,

is:

$$\mathrm{logit}(p_{ijk\ell}) = \alpha_i + \beta\mathbb{I}_j + \delta_k + \gamma$$

(1)

where

the

fixed

ef-

fects

are

GLM for partially pooled categorical pre

$\alpha_i,$

$\beta\mathbb{I}_j$

and

δ_k:

GLM for partially pooled categorical pre

α_i

rep-

re-

sents

the

item

type,

the

in-

di-

ca-

tor

vari-

able

GLM for partially pooled categorical pre

$\mathbb{I}_j$

de-

notes

whether

there

is

cor-

rect

doc-

u-

men-

ta-

tion,

the

co-

ef-

fi-

cient

GLM for partially pooled categorical pre

β

is

the

given

to

the

in-

stance

that

there

is

cor-

rect

doc-

u-

men-

ta-

tion,

and

δ_k

rep-

re-

sents

the

coun-

try

of

provenance.

The

ran-

dom

ef-

fect

γ_ℓ

rep-

re-

sents

the

en-

try

ef-

fect,

which

we

in-

clude

be-

cause

there

may

be

cor-

re-

la-

tions

be-

tween

lines

within

an

en-

try.

We

an-tic-

i-

pate

some

cor-

re-

la-

tion

be-

cause

lines

in

the

same

en-

try

orig-

i-

nate

from

a

com-

mon

con-

text,

SO

they

may

be

non-

compliant

for

re-

lated

rea-

sons.

The

in-

dices

can

take

val-

ues

$i = 1, \ldots, a, \quad a \in \mathbb{N}, \quad a = \#$ i

(2)

$j = 1, 2,$ without

(3)

The

val-

ues

of

the

in-

di-

ca-

tor

vari-

able

are

$$\mathbb{I}_1 = 0, \quad \text{without documentatio}$$

(6)

$$\mathbb{I}_2 = 1, \quad \text{with documentation.}$$

(7)

The

ran-

dom

ef-

fect

γ_ℓ

has

dis-

tri-

bu-

tion

$$\gamma_\ell | \sigma \sim \text{Normal}(0, \sigma), \quad \ell = 1, \ldots$$

(8)

If

all

data

were

in

line

mode

then

the

above

model

would

be

a

fairly

stan-

dard

mixed

ef-

fects

lo-

gistic

re-

gres-

sion

with

cat-

e-

gor-

i-

cal

vari-

ables.

How-

ever,

be-

cause

of

the

use

of

con-

tainer

mode

to

cap-

ture

the

data,

we

don't

observe

out-

comes

for

each

line,

as

ev-

ery

line

in

the

en-

try

is

marked

as

non-

compliant

if

any

line

in

the

en-

try

is

found

to

be

non-

compliant.

There-

fore,

the

out-

come

is

whether

the

en-

try

is

com-

pli-

ant

and

we

need

to

cal-

cu-

late

the

prob-

a-

bil-

ity

that

the

en-

try

is

non-

compliant,

which

is

one

minus

the

prob-

a-

bil-

ity

that

ev-

ery

line

in

the

en-

try

is

com-

pli-

ant:

$$\mathbb{P}\,(\text{Entry } l \text{ non-compliant}) = q_l$$

(9)

where

GLM for partially pooled categorical pre

p_{ijkl}

is

the

prob-

a-

bil-

ity

that

the

line

with

in-

dices

$ijkl$

is

non-

compliant,

cal-

cu-

lated

from

Eq. (**??**).

Hence,

1950*Biometrics, 000* 0000

for

en-

tries

in

con-

tainer

mode,

we

treat

the

en-

try

1962*Biometrics, 000* 0000

as

a

1964*Biometrics, 000* 0000

Bernoulli

ran-

1966*Biometrics, 000* 0000

dom

vari-

able

with

prob-

a-

bil-

ity

de-

fined

1976*Biometrics, 000* 0000

by

Eq. (**??**),

while

for

en-

tries

1982*Biometrics, 000* 0000

in

line

mode,

we

treat

each

line

as

a

Bernoulli

ran-

dom

vari-

able

with

prob-

a-

bil-

ity

as

de-

fined

in

Eq. (**??**).

This

pa-

per

in-

cludes

2010*Biometrics, 000* 0000

three

anal-

y-

ses:

an

asymp-

totic

anal-

y-

sis,

a

sim-

u-

la-

tion

study,

and

a

case

study

of

Aus-

tralian

biose-

cu-

rity

data.

For

the

asymp-

totic

anal-

y-

sis

we

only

con-

sider

the

item

type,

ignor-

ing

ef-

fects

due

to

the

coun-

try

of

origin,

doc-

u-

men-

ta-

tion

and

en-

try

ef-

fect.

As

such,

rather

than

us-

ing

Eq. (**??**),

we

just

con-

sider

the

prob-

a-

bil-

ity

that

a

line

of

item

type

i

is

non-

compliant,

p_i.

The

sim-

u-

la-

tion

study

and

the

case

study

both

use

the

full

model,

as

de-

fined

above.

3. Asymptotic anal- y- sis

We

asymp-

totic

anal-

y-

sis

to

in-

ves-

ti-

gate

how

the

pre-

ci-

sion

of

es-

ti-

mates

de-

pends

on

en-

try

size,

the

num-

ber

of

en-

tries,

the

prob-

a-

bil-

ity

of

non-

compliance

and

whether

item

types

are

mixed.

This

anal-

y-

sis

com-

prises

two

parts.

The

first

as-

sumes

that

all

items

are

a

sin-

gle

type,

which

al-

lows

us

to

quan-

tify

how

the

amount

of

data,

prob-

a-

bil-

ity

of

non-

compliance

and

en-

try

size

af-

fect

pre-

ci-

sion.

The

sec-

ond

part

as-

sumes

that

there

are

two

dif-

fer-

ent

item

types,

and

it

ex-

plores

how

chang-

ing

the

pro-

por-

tion

of

en-

tries

with

both

item

types

mixed

af-

fects

pre-

ci-

sion.

Throughout

this

sec-

tion

we

make

two

sim-

pli-

fi-

ca-

tions.

Firstly,

we

do

not

sep-

a-

rate

line

mode

and

con-

tainer

mode

be-

cause

con-

tainer

mode

data

with

an

en-

try

size

of

one

is

math-

e-

mat-

i-

cally

equiv-

a-

lent

to

line

mode

data.

Hence,

throughout

these

anal-

y-

sis,

an

en-

try

size

of

one

means

line

mode

and

en-

try

size

greater

than

one

im-

plies

con-

tainer

mode.

Sec-

ondly,

we

as-

sume

that

each

item

has

a

fixed

prob-

a-

bil-

ity

of

non-

compliance.

As

such,

any

un-

certainty

in

the

in-

fer-

ence

of

this

value

arises

as

a

re-

sult

of

the

dif-

fer-

ence

be-

tween

con-

tainer

and

line

mode.

With

this

in

hand,

we

con-

sider

each

line

a

Bernoulli

trial,

which

only

de-

pends

on

the

item

type.

When

the

en-

try

size

is

greater

than

one,

the

rel-

e-

vant

prob-

a-

bil-

ity

is

whether

at

least

one

line

was

non-

compliant.

We

es-

ti-

mate

pre-

ci-

sion

us-

ing

an

asymp-

totic

es-

ti-

mate

of

the

stan-

dard

er-

ror.

We

cal-

cu-

late

the

pre-

ci-

sion

from

the

square

roots

of

the

di-

ag-

Onal

el-

e-

ments

in

the

Fisher

in-

for-

ma-

tion

ma-

trix,

$\mathcal{I},$

which

is

the

ex-

pected

value

of

the

neg-

a-

tive

of

the

Hessian

ma-

trix

of

the

log-

likelihood

eval-

u-

ated

at

the

value

of

the

pa-

ram-

e-

ter.

3.1 *Single item type*

For

the

sin-

gle

item

type

case,

we

set

the

prob-

a-

bil-

ity

of

non-

compliance

to

be

$p,$

and

de-

fine

N

as

the

to-

tal

num-

ber

of

en-

tries,

I

as

the

num-

ber

of

non-

compliant

en-

tries

and

GLM for partially pooled categorical pre

S

as

the

size

(i.e.

the

num-

ber

of

lines)

in

each

en-

try.

The

like-

li-

hood

is

a

bi-

no-

mial

dis-

tri-

bu-

tion,

where

the

out-

come

is

the

dis-

COV-

ery

of

a

non-

compliant

en-

try.

The

prob-

a-

bil-

ity

that

an

en-

try

is

com-

pli-

ant

is

$$\mathbb{P}(\text{entry compliant}) = (1-p)^S,$$

0)

meaning

that

the

bi-

no-

mial

like-

li-

hood

for

a

set

en-

try

size

GLM for partially pooled categorical pre

S

pro-

por-

tional

to

$$\mathcal{L}_S = \left(1 - (1-p)^S\right)^I \left(1 - p)^S\right)$$

1)

Therefore,

the

log-

likelihood

is

$$\log \mathcal{L}_S = I \log \left(1 - (1-p)^S\right) +$$

.2)

As

there

is

only

one

pa-

ram-

e-

ter,

we

cal-

cu-

late

its

sec-

ond

deriva-

tive

(rather

than

need-

ing

a

Hessian

ma-

trix):

$$\left[\frac{\partial^2 \log \mathcal{L}_S}{\partial p^2}\right] = \frac{S\left(N + \frac{I((1+S)(1}{((1-p)}\right.}{(1-p)^2}$$

3)

To

cal-

cu-

late

the

Fisher

in-

for-

ma-

tion,

we

take

the

ex-

pec-

ta-

tion

of

the

num-

ber

of

en-

tries

with

non-

compliance,

which

de-

pends

on

the

size

of

the

en-

try:

$$\mathbb{E}[I] = N(1 - (1 - p)^S). \quad (.4)$$

Hence

the

Fisher

in-

for-

ma-

tion

is

$$\mathcal{I} = -\mathbb{E}\left[\frac{\partial^2 \log \mathcal{L}_S}{\partial p^2}\right] = -\frac{NS^2(}{(1-}$$

.5)

and

the

stan-

dard

er-

ror

es-

ti-

mate

is

$$SE = \left(-\frac{NS^2(1-p)^{S-2}}{(1-p)^S - 1}\right)^{-1/2}$$

6)

Using

Eq. (??)

we

can

un-

der-

stand

how

the

prob-

a-

bil-

ity

of

non-

compliance,

the

en-

try

size

and

the

num-

ber

of

lines

af-

fect

the

stan-

dard

er-

ror,

and

we

plot

these

re-

la-

tionships

in

Figure ??.

The

left

plot

shows

that

the

stan-

dard

er-

ror

de-

pends

on

the

prob-

a-

bil-

ity

of

non-

compliance

and

that

the

re-

la-

tionship

de-

pends

on

the

en-

try

size.

For

all

en-

try

sizes,

the

stan-

dard

er-

ror

is

small

when

the

prob-

a-

bil-

ity

of

non-

compliance

is

small

(be-

low

$\sim$

0.3).

How-

ever,

for

larger

val-

ues

of

the

prob-

a-

bil-

ity

of

non-

compliance,

the

stan-

dard

er-

ror

in-

crease

signif-

i-

cantly

if

the

en-

try

size

is

three

or

greater.

The

large

GLM for partially pooled categorical pre

p

be-

haviour

is

driven

by

the

$(1-$

$p)^{S-2}$

term

in

Eq. (**??**),

which

means

SE

goes

to

0

if

$S =$

1,

while

it

di-

verges

if

$S \geq$

3.

Figure ??

also

shows

how

the

stan-

dard

er-

ror

de-

creases

as

the

num-

ber

of

lines

of

data

in-

creases.

The

lower

the

en-

try

size

is,

the

lower

the

stan-

dard

er-

ror,

and,

as

GLM for partially pooled categorical pre

$SE \sim$

$\sqrt{1/N_{E,S}}$,

con-

tainer

mode

data

with

larger

en-

try

sizes

re-

quire

a

large

amount

of

data

to

reach

the

same

stan-

dard

er-

ror.

[Figure 1

about

here.]

3.2 *Two item types*

In

this

sec-

tion

we

con-

sider

a

sit-

u-

a-

tion

where

there

are

two

items

with

prob-

a-

bil-

i-

ties

of

non-

compliance

of

p_1

and

$p_2,$

and

we

ex-

am-

ine

how

these

dif-

fer-

ent

items

prob-

a-

bil-

i-

ties

in-

ter-

act.

We

fo-

cus

on

a

sce-

nario

where

ev-

ery

en-

try

is

of

size

two,

mean-

ing

there

are

three

types

of

en-

tries:

only

type

1;

only

type

2;

or

mixed,

with

one

line

of

type

1

and

one

of

type

2.

We

de-

note

the

num-

ber

of

lines

within

a

sin-

gle

en-

try

of

type

1

and

2

as

S_1

and

GLM for partially pooled categorical pre

S_2

re-

spec-

tively,

and

$I(S_1, S_2$

and

$N(S_1, S_2)$

are

the

num-

ber

of

en-

tries

with

non-

compliance

and

to-

tal

num-

ber

of

en-

tries

with

S_1

type

lines

and

S_2

type

2

lines.

For

our

case,

we

can

have

GLM for partially pooled categorical pre

$S_1 =$

$2, S_2 =$

0;

$$S_1 =$$

$1, S_2 =$

1;

or

$S_1 =$

$$0, S_2 =$$

2.

Rewrit-

ing

the

log-

likelihood

from

Eq. (**??**),

we

get

$$\log \mathcal{L} = \sum_{S_1,S_2} I(S_1, S_2) \log\left(1 - (1-p_1)^{S_1}(1-p_2)^{S_2}\right) +$$

$$(N(S_1, S_2) - I(S_1, S_2)) \log\left((1-p_1)^{S_1}(1-p_2)^{S_2}\right).$$

.7)

We

com-

pute

the

Fisher

in-

for-

ma-

tion

ma-

trix

and

the

stan-

dard

er-

ror

us-

ing

Math-

e-

mat-

ica.

The

stan-

dard

er-

ror

for

p_1

is

$$\frac{1}{2}\sqrt{-\frac{p_1\left(p_1^2-3p_1+2\right)\left(N_{1,2}(p_1\right.}{N_{1,1}(p_1-1)(N_{1,2}(p_1-1}}$$

.8)

where

$N_{1,1}$

and

$N_{2,2}$

are

the

num-

ber

of

en-

tries

with

two

type

1

lines

and

two

type

2

lines

re-

spec-

tively,

while

$N_{1,2}$

are

the

num-

ber

of

en-

tries

with

both

type

1

and

type

2

lines.

The

stan-

dard

er-

ror

for

p_2

is

the

same,

with

$N_{1,1}$

and

$N_{2,2}$

switched

and

p_1

and

p_2

switched.

GLM for partially pooled categorical pre

By

ex-

am-

in-

ing

Eq. (**??**)

we

can

GLM for partially pooled categorical pre

see

that

the

be-

haviour

of

the

stan-

dard

er-

rors

is

more

com-

plex

than

when

we

con-

sid-

ered

only

one

type

of

item.

No-

tably,

the

num-

ber

of

en-

tries

of

only

type

2,

$N_{2,2},$

is

in

the

equa-

tion

for

the

type

1

stan-

dard

er-

ror,

along

with

the

prob-

a-

bil-

ity

of

non-

compliance

of

type

$2,$

$p_2.$

Figure ??

shows

how

the

stan-

dard

er-

ror

varies

with

the

pro-

por-

tion

of

mixed

en-

tries,

for

dif-

fer-

ent

non-

compliance

prob-

a-

bil-

i-

ties.

Here

we

fix

$$N_{\text{total}} =$$

50,

and

keep

$$N_{1,1} =$$

$N_{2,2}$

while

the

pro-

por-

tion

$N_{1,2}/N_{\text{total}}$

is

var-

ied.

In

each

case,

we

hold

p_1

at

0.1,

and

we

vary

p_2

for

0.05

up

to

0.9

across

the

four

plots.

The

way

that

the

stan-

dard

er-

ror

changes

as

a

func-

tion

of

the

pro-

por-

tion

of

mixed

en-

tries

changes

markedly,

de-

pend-

ing

on

the

value

of

$p_2.$

In

partic-

u-

lar,

when

p_2

is

0.7

and

0.9,

the

stan-

dard

er-

ror

for

p_1

in-

creases

as

the

pro-

por-

tion

of

mixed

en-

tries

in-

creases,

while

the

stan-

dard

er-

ror

for

p_2

ac-

tually

de-

creases,

while

the

pro-

por-

tion

of

mixed

en-

tries

is

be-

low

$\sim$

90%.

[Figure 2

about

here.]

3.3 *Asymptotic anal- y- sis con-*

asymp-

totic

anal-

y-

sis

we

make

two

obser-

va-

tions:

GLM for partially pooled categorical pre

(1)

Container

mode

data

are

most

use-ful

when

the

prob-

a-

bil-

ity

of

non-

compliance,

$p,$

is

small.

In

the

ex-

treme

case

where

GLM for partially pooled categorical pre

$$p \rightarrow$$

0,

con-

tainer

mode

data

are

equiv-

a-

lent

to

line

mode

data

be-

cause

when-ever

an

en-

try

is

found

to

be

com-

pli-

ant,

we

know

that

ev-

ery

line

within

that

en-

try

is

com-

pli-

ant,

re-

gard-

less

of

whether

it

is

con-

tainer

mode

or

line

mode.

(2)

By

al-

low-

ing

en-

tries

to

con-

tain

lines

of

dif-

fer-

ent

item

types,

con-

tainer

mode

can

in-

crease

or

de-

crease

the

pre-

ci-

sion

of

es-

ti-

mates,

de-

pend-

ing

on

the

true

prob-

a-

bil-

ity

of

non-

compliance.

4. Simulation study

As

the

first

step

in

de-

vel-

op-

ing

a

method

to

anal-

yse

real-

world

data

sets,

we

sim-

u-

late

a

data

set

that

con-

tains

key

com-

plex-

i-

ties

in-

clud-

ing

mul-

ti-

ple

item

types,

fixed

ef-

fects

and

ran-

dom

ef-

fects.

Once

we

have

sim-

u-

lated

data,

we

can

fit

a

model

to

es-

ti-

mate

pa-

ram-

e-

ters.

The

advantage

of

our

ap-

proach

is

that

we

can

GLM for partially pooled categorical pre

(1)

ver-

ify

that

our

model

be-

haves

cor-

rectly

and

(2)

ex-

plore

how

model

pre-

ci-

sion

varies

in

a

more

com-

plex

set-

ting.

4.1 *Data*

sim-

u-

la-

tion

u-

late

data

from

the

full

model

(as

de-

scribed

in

Sec-

tion **??**).

For

our

sim-

u-

la-

tions,

we

chose

pa-

ram-

e-

ters

that

we

ex-

pect

to

be

sim-

i-

lar

to

real-

world

val-

ues.

We

set

GLM for partially pooled categorical pre

$\beta =$

$-1,$

choos-

ing

an

neg-

a-

tive

ef-

fect

of

hav-

ing

cor-

rect

doc-

u-

men-

ta-

tion,

be-

cause

we

ex-

pect

lines

with

cor-

rect

doc-

u-

men-

ta-

tion

have

a

higher

chance

of

be-

ing

com-

pli-

ant.

We

in-

clude

$a =$

5

item

types

with

GLM for partially pooled categorical pre

α_i

tak-

ing

val-

ues

of

–

6.91,

–

4.60,

–

3.89,

–

2.94,

–

1.39

(cor-

re-

spond-

ing

to

non-

compliance

prob-

a-

bil-

i-

ties

of

0.001,

$0.01,$

0.02,

0.05

and

$0.2,$

if

all

other

ef-

fects

were

0).

We

chose

neg-

a-

tive

val-

ues

for

the

item

ef-

fect

GLM for partially pooled categorical pre

$\alpha_i,$

since

the

prob-

a-

bil-

ity

of

de-

tec-

tion

is

ex-

pected

to

be

low

for

any

item

in

real-

world

data.

We

used

GLM for partially pooled categorical pre

$d =$

3

coun-

tries

and

set

their

weights

to

be

$0.5,$

–

1

and

0.25.

For

within-entry

cor-

re-

la-

tion,

we

set

GLM for partially pooled categorical pre

$\sigma =$

0.25.

For

each

en-

try

we

draw

whether

it

is

in

line

mode

or

not

with

prob-

a-

bil-

ity

0.25,

and

for

each

line

we

set

the

prob-

a-

bil-

ity

of

hav-

ing

cor-

rect

doc-

u-

men-

ta-

tion

to

be

0.2.

We

show

an

ex-

am-

ple

of

the

for-

mat

of

the

data

in

Ta-

ble

??.

[Table 1

about

here.]

4.2 *STAN*

model

We

model

the

sys-

tem

in

a

Bayesian

framework,

us-

ing

the

RSTAN

pack-

age

in

R.

We

choose

a

Bayesian

pack-

age

due

to

the

ease

of

spec-

i-

fy-

ing

the

model

and

could

be

updated,

in

prin-

ci-

ple,

for

any

spec-

i-

fi-

ca-

tion

of

the

model

for

the

prob-

a-

bil-

ity

of

non-

compliance.

The

basic

struc-

ture

for

fit-

ting

the

sim-

u-

lated

data

is

to

de-

fine

the

prob-

a-

bil-

ity

of

non-

compliance

for

each

line

(fol-

low-

ing

Eq. (**??**))

$$\mathrm{logit}(p_{ijk\ell}) = \alpha_i + \beta\mathbb{I}_j + \delta_k + \gamma$$

.9)

where

$$\gamma_\ell | \sigma \sim \mathrm{Normal}(0, \sigma), \quad \ell = 1, \ldots$$

20)

As

we

fit

our

model

in

a

Bayesian

framework,

we

set

priors

on

each

pa-

ram-

e-

ter:

$$\alpha_i \sim \text{Normal}(-4, 4), \quad i = 1, . .$$

21)

$$\beta \sim \text{Normal}(0, 0.5),$$

22)

The

prob-

a-

bil-

ity

$p_{ijk\ell}$

is

the

prob-

a-

bil-

ity

of

non-

compliance

of

a

line

that

has

char-

ac-

ter-

is-

tics

(i, j, k, ℓ).

We

de-

note

the

prob-

a-

bil-

ity

that

line

GLM for partially pooled categorical pre

n

is

non-

compliant

as

$p_{(n)}$,

where

$$p_{(n)} =$$

GLM for partially pooled categorical pre

$p_{ijk\ell}$

if

that

line

has

char-

ac-

ter-

is-

tics

(i, j, k, ℓ).

We

fit

the

STAN

model

to

the

sim-

u-

lated

data

for

two

rea-

sons.

Firstly

we

con-

firm

that,

with

suf-

fi-

cient

data,

the

STAN

model

can

ac-

cu-

rately

es-

ti-

mate

model

pa-

ram-

e-

ters.

Sec-

ondly

we

ex-

plore

how

chang-

ing

the

amount

of

data

and

the

pro-

por-

tion

of

con-

tainer

mode

data

af-

fects

pre-

ci-

sion,

over

a

range

of

en-

try

sizes.

To

measure

pre-

ci-

sion,

we

use

the

stan-

dard

de-

a-

tion

of

the

pos-

terior

sam-

ples,

be-

cause

it

is

the

Bayesian

equiv-

a-

lent

of

the

stan-

dard

er-

ror

cal-

cu-

la-

tions

in

Sec-

tion ??.

4.3 *Simulation re- sults*

Figure ?? shows

the

that

the

STAN

model

gives

good

pa-

ram-

e-

ter

es-

ti-

mates,

if

there

is

suf-

fi-

cient

data.

Over-all,

the

STAN

es-

ti-

mates

are

close

to

the

true

val-

ues

when

there

is

a

large

amount

of

data.

The

two

small-est

val-

ues

of

GLM for partially pooled categorical pre

α

show

the

worst

per-

for-

mance,

which

is

not

surprising,

given

our

asymp-

totic

anal-

y-

sis

re-

sults

(Sec-

tion ??).

[Figure 3

about

here.]

The

pre-

ci-

sion

of

model

es-

ti-

mates

de-

pends

on

the

en-

try

size,

the

prob-

a-

bil-

ity

of

non-

compliance

and

the

amount

of

con-

tainer

mode

data

(Figure **??**).

The

best-

case

sce-

nario

is

when

all

data

are

in

line

mode

(ratio

of

con-

tainer

mode

$=$

$0),$

and

the

dis-

tance

from

that

line

shows

the

im-

pact

of

us-

ing

a

mix,

or

only,

con-

tainer

mode

data.

We

find

that

the

dif-

fer-

ence

be-

tween

the

pre-

ci-

sion

es-

ti-

mates

de-

pends

strongly

on

the

com-

bi-

na-

tion

of

factors.

There

are

some

com-

bi-

na-

tions

where

all

anal-

y-

ses

re-

turn

sim-

i-

lar

pre-

ci-

sion

(e.g.

GLM for partially pooled categorical pre

$\alpha =$

-1.4

and

en-

try

size

10),

while

other

com-

bi-

na-

tions

have

a

large

gap

where

the

line

only

data

far

out-

per-

forms

the

oth-

ers

(e.g.

GLM for partially pooled categorical pre

$\alpha =$

-4.6

and

en-

try

size

$=$

5).

[Figure 4

about

here.]

5. Case study

We

ap-

ply

our

model

to

some

real

biose-

cu-

rity

data,

to

GLM for partially pooled categorical pre

see

how

our

un-

der-

stand-

ing

of

the

sys-

tem

changes

GLM for partially pooled categorical pre

by

fit-

ting

the

full

data,

com-

pared

to

only

us-

ing

the

con-

tainer

mode

data.

The

data

set

is

of

furni-

ture

im-

ports

in

GLM for partially pooled categorical pre

2020,

and

we

break

the

data

set

into

weekly

data

sets

for

this

anal-

y-

sis.

Be-

cause

con-

tainer

mode

data

with

no

non-

compliance

is

equiv-

a-

lent

to

line

mode

data,

the

more

container-

mode

en-

tries

with

non-

compliance,

the

larger

the

dif-

fer-

ences

could

be

be-

tween

analysing

all

of

the

data

com-

pared

to

only

the

line

mode

data.

There

is

a

nat-

u-

ral

break

in

the

dataset,

where

three

weeks

have

four

con-

tainer

mode

en-

tries

with

non-

compliance,

with

the

re-

main-

ing

weeks

hav-

ing

fewer.

Hence,

we

choose

to

anal-

yse

these

three

weeks

as

a

demonstra-

tion

of

po-

ten-

tial

real-

world

dif-

fer-

ences

be-

tween

analysing

all

data

and

only

line

mode

data.

For

con-

fi-

dential-ity,

for

each

week

we

re-

name

the

item

type,

the

coun-

try

and

the

week

to

be

in-

tegers

(week

1,

2

and

3

do

not

cor-

re-

spond

to

the

first

3

weeks

of

the

year,

and

coun-

try

1

and

item

1

are

not

the

same

in

weeks

1

and

2).

Re-

la-

bel-

ing

the

data

does

not

cause

is-

sues

be-

cause

we

are

not

try-

ing

to

com-

pare

es-

ti-

mates

be-

tween

weeks

or

draw

in-

fer-

ences

be-

tween

coun-

tries

or

item

types

in

this

anal-

y-

sis.

A

summary

of

the

data

is

given

in

Ta-

ble ??.

GLM for partially pooled categorical pre

[Table 2

about

here.]

We

fit

our

STAN

model

to

the

data

from

each

week,

us-

ing

all

the

data

and

the

line-

only

sub-

set

of

the

data.

Be-

cause

the

line-

only

anal-

GLM for partially pooled categorical pre

y-

sis

is

a

sub-

set

of

the

data,

not

ev-

ery

coun-

try

or

item

is

present

in

the

line-

only

data

set.

For

the

pa-

ram-

e-

ters

that

do

match,

we

gen-

er-

ate

a

scat-

ter

plot

of

the

mean

es-

ti-

mates

to

GLM for partially pooled categorical pre

see

how

frequently

we

get

dif-

fer-

ent

es-

ti-

mates

(Figure **??**).

For

many

of

the

pa-

ram-

e-

ters,

we

get

very

sim-

i-

lar

re-

sults

ir-

re-

spec-

tive

of

which

data

we

use.

This

is

not

surprising,

given

that

most

of

the

data

are

in

line

mode

and

that

line

mode

data

gives

more

in-

for-

ma-

tion

than

con-

tainer

mode

data.

How-

ever,

it

is

in-

ter-

est-

ing

that

in-

clud-

ing

the

con-

tainer

mode

data

re-

sults

in

quite

large

changes

to

some

of

the

pa-

ram-

e-

ter

es-

ti-

mates.

[Figure 5

about

here.]

Analysing

the

full

data

set,

rather

than

the

line-

only

data,

gives

us

in-

for-

ma-

tion

about

more

items

and

coun-

tries.

Ta-

ble ??

shows

all

of

the

pa-

ram-

e-

ter

es-

ti-

mates

that

we

only

get

when

in-

clud-

ing

the

con-

tainer

mode

data

in

the

anal-

y-

sis.

[Table 3

about

here.]

6. Discussion

In

this

pa-

per

we

sought

to

un-

der-

stand

how

con-

tainer

mode

data

af-

fects

our

abil-

ity

to

es-

ti-

mate

risks

in

the

biose-

cu-

rity

set-ting.

Due

to

the

rel-

a-

tive

ease

of

analysing

line

data,

in

our

ex-

perience,

con-

tainer

mode

data

are

of-

ten

ex-

cluded

from

data

anal-

y-

sis.

From

our

anal-

y-

sis

of

real

biose-

cu-

rity

data,

we

find

that

in-

clud-

ing

the

con-

tainer

mode

data

in

anal-

y-

sis

can

markedly

change

some

re-

sults

and

that

by

only

us-

ing

line

data

the

anal-

y-

sis

is

some-what

limited

there

are

pa-

ram-

e-

ter

es-

ti-

mates

that

sim-

ply

can-

not

be

made.

Through

our

sim-

u-

la-

tion

study

and

our

asymp-

totic

anal-

y-

sis

we

gain

an

un-

der-

stand-

ing

of

how

mix-

ing

con-

tainer

mode

data

into

line

mode

data

im-

pacts

our

abil-

ity

to

make

pre-

cise

in-

fer-

ences.

We

GLM for partially pooled categorical pre

see

that

mix-

ing

lines

that

have

dif-

fer-

ent

prob-

a-

bil-

i-

ties

of

be-

ing

non-

compliant

has

di-

ver-

gent

af-

fects

on

model

pre-

ci-

sion.

In

the

asymp-

totic

anal-

y-

sis

we

see

that

mix-

ing

a

high-

probability

item

with

a

lower

prob-

a-

bil-

ity

item

makes

it

easier

to

pre-

cisely

es-

ti-

mate

the

high-

probability

item's

pa-

ram-

e-

ter,

while

mix-

ing

two

low-

probability

items

makes

it

chal-

leng-

ing

to

make

in-

fer-

ences

about

ei-

ther.

When

one

item

has

a

low

prob-

a-

bil-

ity

of

non-

compliance,

a

non-

compliant

mixed

en-

try

is

likely

to

cor-

re-

spond

to

a

non-

compliant

en-

try

of

the

of

the

item

with

a

higher

prob-

a-

bil-

ity

of

com-

pli-

ance.

The

ex-

tra

data

from

this

en-

try

can

there-

fore

be

used

to

re-

fine

the

pa-

ram-

e-

ter

es-

ti-

mates,

whereas

if

the

re-

spec-

tive

prob-

a-

bil-

i-

ties

are

not

well

re-

solved

this

is

not

pos-

si-

ble.

These

in-

ter-

ac-

tions

be-

tween

items

are

ex-

ac-

er-

bated

in

the

sim-

u-

la-

tion

study,

where

we

have

larger

en-

tries

and

more

item

types,

and

we

GLM for partially pooled categorical pre

see

dif-

fer-

ing

patterns

of

item

types,

en-

try

sizes

and

amounts

of

data

where

the

pres-

ence

of

con-

tainer

mode

data

de-

grades

model

pre-

ci-

sion,

com-

pared

to

a

line-

only

anal-

y-

sis.

While

we

are

able

to

model

the

full

dataset,

the

con-

tin-

u-

ing

pres-

ence

of

con-

tainer

mode

data

will

be

a

barrier

for

fu-

ture

data

anal-

y-

sis.

Most

sta-

tis-

ti-

cal

and

ma-

chine

learn-

ing

al-

gorithms

are

de-

signed

such

that

each

row

of

data

will

have

ex-

planatory

vari-

ables

and

an

out-

come.

With

con-

tainer

mode

data,

we

have

the

ex-

planatory

vari-

ables,

but

only

partial

in-

for-

ma-

tion

about

the

out-

come.

Hence,

we

need

customised

al-

gorithms

(such

as

de-

scribed

in

this

pa-

per)

to

anal-

yse

it

ap-

pro-

priately.

Even

so,

many

sim-

pli-

fy-

ing

as-

sump-

tions

have

been

made:

for

ex-

am-

ple,

that

a

given

item's

prob-

a-

bil-

ity

of

non-

compliance

does

not

vary

over

time.

These

sim-

pli-

fi-

ca-

tions

do

not

change

our

con-

clusions,

as

the

anal-

y-

sis

still

serves

to

high-

light

the

dif-

fer-

ences

be-

tween

con-

tainer

and

line

mode.

How-

ever,

without

proper

care,

the

specific

re-

sults

of

our

case-

study

anal-

y-

sis

should

not

be

ap-

plied

to

de-

ci-

sion

mak-

ing.

While

in

our

case

the

pool-

ing

of

out-

comes

in

con-

tainer

mode

is

an

ar-

ti-

fact

of

the

sys-

tem,

pooled

test-

ing

are

of-

ten

de-

signed

strategically

in

public

health

to

im-

prove

ef-

fi-

ciency.

The

strain

on

the

PCR

test-

ing

sys-

tem

has

prompted

regimes

to

iden-

tify

cases

while

minimis-

ing

the

rounds

of

test-

ing (**?**).

While

it

is

clear

that

the

way

pool-

ing

is

con-

ducted

af-

fects

how

quickly

in-

fec-

tions

can

be

iden-

ti-

fied,

in

some

circumstances,

the

out-

comes

of

pooled

test-

ing

can

also

be

used

to

es-

ti-

mate

pa-

ram-

e-

ters,

and

our

work

has

im-

pli-

ca-

tions

for

these

sit-

u-

a-

tions

(????).

From

our

asymp-

totic

anal-

GLM for partially pooled categorical pre

y-

sis,

it

is

clear

that

even

when

only

es-

ti-

mat-

ing

a

preva-

lence,

for

a

fixed

num-

ber

of

tests,

the

stan-

dard

er-

ror

of

the

es-

ti-

mate

de-

pends

on

pool

size

and

there-

fore

there

would

be

an

op-

ti-

mal

pool

size,

which

would

de-

pend

on

the

preva-

lence

(?).

When

there

is

a

model

with

pa-

ram-

e-

ters

be-

ing

es-

ti-

mated,

the

added

com-

plex-

ity

could

com-

pound

the

is-

sue,

mak-

ing

pool

siz-

ing

a

more

im-

por-

tant

as-

pect

of

study

de-

sign.

Further-more,

as

we

demonstrate

here,

it

is

not

only

pool

size

that

mat-

ters,

it

is

how

sam-

ples

with

dif-

fer-

ent

char-

ac-

ter-

is-

tics

are

grouped

to-

gether.

Hence,

for

a

given

study,

there

may

not

be

one

op-

ti-

mal

pool-

ing

strategy,

and

there

would

likely

be

tradeoffs

be-

tween

dif-

fer-

ent

pa-

ram-

e-

ters

when

it

comes

to

maximis-

ing

the

pre-

ci-

sion

of

es-

ti-

mates.

In

this

pa-

per

we

have

in-

ves-

ti-

gated

how

pooled

data

can

re-

duce

our

abil-

ity

to

make

sta-

tis-

ti-

cal

in-

fer-

ences

about

the

pop-

u-

la-

tion.

Within

the

Aus-

tralian

biose-

cu-

rity

con-

text,

the

pool-

ing

is

due

to

how

data

are

recorded

and

was

im-

plemented

for

op-

er-

a-

tional

rea-

sons.

Due

to

this

op-

er-

a-

tional

de-

ci-

sion,

analysing

full

datasets

is

harder

than

the

line-

only

mode

data,

mean-

ing

that

ei-

ther

anal-

y-

sis

is

re-

stricted

to

a

sub-

set

of

pos-

si-

ble

methods

or

that

con-

tainer

mode

data

is

ignored.

While

it

does

not

fol-

low

that

con-

tainer

mode

should

be

re-

moved

due

to

op-

er-

a-

tional

ef-

fi-

ciency,

if

the

line-

level

data

could

be

cap-

tured

then

it

would

im-

prove

our

un-

der-

stand-

ing

of

biose-

cu-

rity

risk.

How-

ever,

pooled

data

will

con-

tinue

to

be

col-

lected

in

var-

i-

ous

fields,

for

ex-

am-

ple

due

to

the

costsavings

of

test-

ing

mul-

ti-

ple

sam-

ples

at

once

for

dis-

eases.

While

there

is

ex-

ten-

sive

work

in

pooled

test-

ing

pro-

to-

cols

for

case

iden-

ti-

fi-

ca-

tion,

there

is

less

work

on

iden-

ti-

fy-

ing

pool

sizes

when

aim-

ing

to

make

in-

fer-

ences

about

as-

pects

of

the

pop-

u-

la-

tion.

Our

work

shows

that

pool-

ing

tests

with

vari-

able

un-

der-

ly-

ing

preva-

lence

af-

fects

pre-

ci-

sion

dif-

fer-

en-

tially.

Hence,

care-ful

con-

sid-

er-

ing

should

be

given

to

de-

sign-

ing

pools

when

the

data

will

be

used

to

make

in-

fer-

ences

about

the

pop-

u-

la-

tion.

7. Code availability

All

is

avail-

able

at

https:

//

`github.`

com/

GLM for partially pooled categorical pre

`cmbaker00/`

`container-line-analysis-p`

8. Data availability

The

that

sup-

port

the

findings

of

this

sim-

u-

la-

tion

study

are

openly

avail-

able

at

https:

//

github.

com/

GLM for partially pooled categorical pre

`cmbaker00/`

container-line-analysis-p

The

case

study

data

are

avail-

able

from

the

Aus-

tralian

De-

part-

ment

of

Agri-

cul-

ture,

Fisheries

and

Forestry.

Re-

stric-

tions

ap-

ply

to

the

avail-

abil-

ity

of

these

data,

which

were

used

un-

der

li-

cense

for

this

study.

9. Conflict of interest

thors

de-

clare

no

con-

flict

of

in-

ter-

est.

Acknowledgements

We

would

like

to

thank

Tom

War-

ing

for

their

con-

tri-

bu-

tion

to

edit-

ing

this

manuscript.

Figure 1: The stan- dard er- ror (Eq. (**??**))

Figure 2: The stan- dard

Figure 3: Re-sults from the STAN model, fit-

Figure
4:
The
stan-
dard
de-
vi-
a-
tion

Figure 5: Scat- ter plot of pa- ram- e-

Table 1: Ex- am- ple sim- u- lated data.

Table
2:
The
num-
ber
of
lines
in
con-

Table 3: Parameter estimates