

Documentation for PROC DQMATCH Enhancements in Release 9.1.3

- [Introduction](#)
- [Overview](#)
- [Syntax](#)
- [CONDITION=](#)
- [exact](#)
- [Example #1](#)
- [Example #2](#)
- [Example #3](#)
- [Notes](#)
- [History](#)

[Introduction](#)

This contains documentation to enhancements in PROC DQMATCH in 9.1.3, but the documentation was not able to get in to the SAS Online Help in time. Therefore, the documentation is being made available here.

This is a preliminary draft of this information.

This information will be added to the SAS Online Help in 9.2.

[Back to Top of Page](#)

Overview

In Release 9.1.3 of SAS Data Quality Server two enhancements were made to the DQMATCH procedure:

- The ability to group criteria together into conditions which are then OR'ed together.
- The ability to use the exact value of a variable when doing clustering.

The enhancements can be used in conjunction with one another, but they don't have to be.

[Back to Top of Page](#)

Syntax

The complete syntax for the CRITERIA statement is now as follows:

```
CRITERIA DELIMSTR=delimited-input-variable | VAR=input-variable
MATCHDEF=match-definition | EXACT
<SENSITIVITY=complexity-value>
<MATCHCODE=output-variable>
<CONDITION=int>
;
```

[Back to Top of Page](#)

CONDITION=

The addition of the CONDITION= option on the criteria statement provides a powerful tool for doing more advanced clustering by allowing you to group criteria together into conditions. For example, you can define a PROC DQMATCH step which will cluster observations where an individual's name and address match, or where the organization's name and address match. The most common use for this functionality is doing householding.

For example, these three records will be a part of the same cluster if you define your criteria such that the name and address match or the organization and address match:

NAME	COMPANY	ADDRESS
Mr. Robert Smith	DataFlux Corporation	4001 Weston Pkwy
	The DataFlux Corp.	4001 Westin Pkwy
Bob Smith		4001 Weston Parkway

Additional details:

- The CONDITION= option accepts a single integer as an argument. Multiple criteria within the same condition are AND'ed together, and multiple conditions are OR'ed together. CONDITION= is optional, and if not specified then that criteria defaults to CONDITION=1.
- If more than one condition is defined it is assumed that you want to perform clustering, so a variable named CLUSTER will be created for you automatically. If you have not specified CLUSTER= on the PROC DQMATCH statement, then it will be created with the default variable name of CLUSTER.
- If more than one condition is defined, specifying MATCHCODE= on the PROC DQMATCH statement will give an error. If only one condition is defined, then this will continue to create a composite match code for that one condition.
- There is a limit of 10 criteria in a single condition.
- There is a limit of 10 conditions.
- There is a limit of 10 criteria statements that specify EXACT (over the entire step, not per condition).

[Back to Top of Page](#)

EXACT

The addition of the EXACT option on the criteria statement allows you to use the exact value of a variable when doing clustering rather than using the match code for that value.

Additional details:

- Since no match definition is being used to interpret the value, you cannot specify MATCHDEF=, MATCHCODE=, or SENSITIVITY= on any CRITERIA statement that specifies EXACT.
- If one or more exact criteria are defined you can still specify MATCHCODE= on the PROC DQMATCH statement, as long as only one condition is defined, but only the non-exact criteria will be represented in the resulting composite match code.
- You cannot specify MATCHCODE= on the PROC DQMATCH statement if all of the CRITERIA statements specify EXACT.
- If EXACT is specified on any CRITERIA statement, it is assumed that you want to perform clustering, so a variable named CLUSTER will be created for you automatically. If you have not specified CLUSTER= on the PROC DQMATCH statement, then it will be created with the default variable name of CLUSTER.
- If EXACT is specified on any of the CRITERIA statements, a composite match code will not be created automatically. You must specify MATCHCODE= on the PROC DQMATCH statement to receive it.

[Back to Top of Page](#)

Example #1

This is an example of fairly typical clustering using two conditions. Just as you do for PROC DQMATCH with only one condition, you must take care when defining your criteria and conditions to get the sort of clustering result that is acceptable for your application.

```
/* Load the ENUSA qkb. This assumes that the system option
   DQSETUPLOC is already set. */
%dqload(dqlocale=(enusa))

/* Define some sample data. */
data customer;
  length custid 8 name org addr $ 20;
  input custid name $char20. org $char20. addr $char20.;
cards;
1 Mr. Robert Smith      DataFlux Corporation 4001 Weston Blvd.
2                        The DataFlux Corp.   4001 Westin Ave
3 Bob Smith              4001 Weston Parkway
4 Sandi Baker            SAS Institute         123 N Main Street
5 Mrs. Sandra Baker     SAS Inst.             401 Oak Ave.
6 sandie smith Baker    DataFlux Corp.       123 Maine Street
7 Bobby J. Smythe       ABC Plumbing         4001 Weston Pkwy
```

```

;
run;

/* Generate the cluster data.  Since more than one condition
   is defined, a variable named CLUSTER will automatically
   be created. */
proc dqmatch data=customer
              out=customer_out;
  criteria condition=1 var=name sensitivity=85 matchdef='Name';
  criteria condition=1 var=addr sensitivity=70 matchdef='Address';

  criteria condition=2 var=org sensitivity=85
matchdef='Organization';
  criteria condition=2 var=addr sensitivity=70 matchdef='Address';
run;

/* Print the result. */
proc print data=customer_out noobs;
run;

```

The result is as follows:

	custid	name	org	addr
CLUSTER				
1	4	Sandi Baker	SAS Institute	123 N Main Street
1	6	sandie smith Baker	DataFlux Corp.	123 Maine Street
2	1	Mr. Robert Smith	DataFlux Corporation	4001 Weston Blvd.
2	7	Bobby J. Smythe	ABC Plumbing	4001 Weston Pkwy
	3	Bob Smith		4001 Weston
	2		The DataFlux Corp.	4001 Westin Ave
	5	Mrs. Sandra Baker	SAS Inst.	401 Oak Ave.
.				

[Back to Top of Page](#)

Example #2

This example uses the EXACT option, but it only uses one condition (since condition is not specified, both criteria default to CONDITION=1). In this case we want to do clustering when the name at a low sensitivity and the exact customer id match. Although we only have one condition, a cluster variable is automatically generated since we specified at least one exact criteria. We have also chosen to generate a match code for the non-exact criteria.

```

/* Load the ENUSA qkb.  This assumes that the system option
   DQSETUPLOC is already set. */

```

```

%dqload(dqlocale=(enusa))

/* Define some sample data. */
data customer;
  length custid $ 5 name addr $ 20;
  input custid $char5. name $char20. addr $char20.;
cards;
00001 Mr. Robert Smith      101 astor way
00001                               101 astor way
00001 Bob Smith             101 S. Astor Wy
00001 Tom Jones             101 Aster Way
00001 Susan Taylor          101 Aster Way
00001 Cindy Taylor          345 main st
00001 Cindy J. Taylor       987 hilltop lane
00002 Sandi Baker
00002 Mike Baker            302 Oxbow Ct.
00002 Mrs. Sandra Baker     117 Windgate Ln
00002 sandie smith Baker    117 Wingate Lane
00002 S. Baker              117 Wind Gate Ct.
00003 S. Baker
;
run;

/* Generate the cluster data. */
proc dqmatch data=customer
  out=customer_out;
  criteria var=custid exact;
  criteria var=name sensitivity=50 matchdef='Name';
run;

/* Print the result. */
proc print data=customer_out noobs;
run;

```

The result is as follows:

custid	name	addr	CLUSTER
00001		101 astor way	.
00001	Mr. Robert Smith	101 astor way	1
00001	Bob Smith	101 S. Astor Wy	1
00003	S. Baker		.
00002	Sandi Baker		2
00002	S. Baker	117 Wind Gate Ct.	2
00002	sandie smith Baker	117 Wingate Lane	2
00002	Mrs. Sandra Baker	117 Windgate Ln	2
00002	Mike Baker	302 Oxbow Ct.	2
00001	Susan Taylor	101 Aster Way	3
00001	Cindy J. Taylor	987 hilltop lane	3
00001	Cindy Taylor	345 main st	3
00001	Tom Jones	101 Aster Way	.

[Back to Top of Page](#)

Example #3

This example combines the use of the two new enhancements by clustering observations where the name and the exact customer id match, or where the address and the exact customer id match.

```

/* Load the ENUSA qkb. This assumes that the system option
   DQSETUPLOC is already set. */
%dqload(dqlocale=(enusa))

/* Define some sample data. */
data customer;
  length custid $ 5 name addr $ 20;
  input custid $char5. name $char20. addr $char20.;
cards;
00001 Mr. Robert Smith      101 astor way
00001                               101 astor way
00001 Bob Smith            101 S. Astor Wy
00001 Tom Jones            101 Aster Way
00001 Susan Taylor         101 Aster Way
00001 Cindy Taylor        345 main st
00001 Cindy J. Taylor      987 hilltop lane
00002 Sandi Baker
00002 Mike Baker           302 Oxbow Ct.
00002 Mrs. Sandra Baker    117 Windgate Ln
00002 sandie smith Baker  117 Wingate Lane
00002 S. Baker             117 Wind Gate Ct.
00003 S. Baker
;
run;

/* Generate the cluster data. */
proc dqmatch data=customer
  out=customer_out;
  criteria condition=1 var=custid exact;
  criteria condition=1 var=name sensitivity=50 matchdef='Name';

  criteria condition=2 var=custid exact;
  criteria condition=2 var=addr sensitivity=70 matchdef='Address';
run;

/* Print the result. */
proc print data=customer_out noobs;
run;

```

The result is as follows:

custid	name	addr	CLUSTER
00001		101 astor way	1
00001	Mr. Robert Smith	101 astor way	1
00001	Susan Taylor	101 Aster Way	1
00001	Tom Jones	101 Aster Way	1
00001	Bob Smith	101 S. Astor Wy	1
00001	Cindy J. Taylor	987 hilltop lane	1
00001	Cindy Taylor	345 main st	1
00003	S. Baker		.
00002	Sandi Baker		2
00002	S. Baker	117 Wind Gate Ct.	2

00002	sandie smith Baker	117 Wingate Lane	2
00002	Mrs. Sandra Baker	117 Windgate Ln	2
00002	Mike Baker	302 Oxbow Ct.	2

One of the important things to notice with this example is how the observations “chain” together. For example, compare the “Mr. Robert Smith” observation and the “Cindy Taylor” observation, which are both placed into cluster #1. At first glance they don’t appear to have anything in common except for the CUSTID=00001. But the observation with “Mr. Robert Smith” matches based on the ADDR/CUSTID with the “Susan Taylor” record, which in turn matches based on the NAME/CUSTID with the “Cindy Taylor” record. This is often exactly the advantage of using multiple conditions, however it also demonstrates why you have to be very careful in defining your conditions and criteria to give the type of result that is appropriate for your application.

[Back to Top of Page](#)

Notes

Copyright 2004, SAS Institute Inc., Cary, NC, USA.

All rights reserved. Produced in the United States of America.

**This material is a draft and SAS Institute makes no warranty about the accuracy of this document. The company is not responsible for errors in the text or for any discrepancies between this documentation and the product described as it now exists or will exist.

SAS® and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration.

[Back to Top of Page](#)